MÉTODOS QUANTITATIVOS COM STATA®


CÓPIA NÃO AUTORIZADA É CRIME
ABDR
ASSOCIAÇÃO BRASILEIRA DE DIREITOS REPROGRÁFICOS
RESPEITE O DIREITO AUTORAL

MÉTODOS QUANTITATIVOS COM STATA®

1ª EDIÇÃO

LUIZ PAULO FÁVERO (ORG.)
PATRÍCIA BELFIORE
RENATA TUROLA TAKAMATSU
JANILSON SUZART

Copidesque: Edna da Silva Cavalcanti
Editoração Eletrônica: Thomson Digital
Revisão Gráfica: Lara Alves

Elsevier Editora Ltda.
Conhecimento sem Fronteiras

Rua Sete de Setembro, 111 – 16° andar
20050-006 – Centro – Rio de Janeiro – RJ – Brasil

Rua Quintana, 753 – 8° andar
04569-011 – Brooklin – São Paulo – SP

Serviço de Atendimento ao Cliente
0800-0265340 sac@elsevier.com.br

ISBN: 978-85-352-5157-9
ISBN (versão eletrônica): 978-85-352-5158-6

Nota: Muito zelo e técnica foram empregados na edição desta obra. No entanto, podem ocorrer erros de digitação, impressão ou dúvida conceitual. Em qualquer das hipóteses, solicitamos a comunicação ao nosso Serviço de Atendimento ao Cliente, para que possamos esclarecer ou encaminhar a questão.

Nem a editora nem o autor assumem qualquer responsabilidade por eventuais danos ou perdas a pessoas ou bens, originados do uso desta publicação.

CIP-BRASIL. CATALOGAÇÃO-NA-FONTE
SINDICATO NACIONAL DOS EDITORES DE LIVROS, RJ

M552
Métodos quantitativos com stata : procedimentos, rotinas e análise de resultados / Luiz Paulo Fávero ... [et al.]. - 1. ed. - Rio de Janeiro : Elsevier, 2014.
23 cm.
ISBN 978-85-352-5157-9
1. Tecnologia da informação. 2. Sistemas operacionais (Computadores). 3. Computadores. 4. Informática. 5. Software. 6. Computadores - Equipamento de entrada e saída. I. Fávero, Luiz Paulo. II. Título.

13-03450 CDD: 004
CDU: 004

APRESENTAÇÃO

Este livro pode ser considerado resultado de várias discussões e elucubrações, ao longo dos últimos anos, sobre a importância da modelagem aplicada aos mais diversos campos do conhecimento humano. O crescente acúmulo de dados gerados, cada vez com maior frequência, em ambientes acadêmicos e organizacionais vem acompanhado do profundo desenvolvimento computacional e do aprimoramento dos softwares estatísticos e econométricos. Dentro deste contexto, o Stata® é um software com grande capacidade de processamento de enormes bases de dados, além de ser capaz de elaborar os mais diversos testes e modelos apropriados e robustos a cada situação e de acordo com aquilo que o pesquisador e o tomador de decisão desejam.

O software Stata® surgiu em 1985. Sua primeira versão, criada por William Gold, era compatível com o sistema operacional DOS. Atualmente, na versão 12, é distribuído e utilizado em mais de 150 países, sendo compatível, por meio do programa Stat/Transfer, com a grande maioria dos softwares que utilizam bases de dados, como Excel, SPSS, SAS, FoxPro, Gauss, LIMDEP, Matlab, Minitab, R, S-PLUS, Statistica, entre outros.

Além disso, o Stata® propicia ao usuário utilizar menus automáticos do tipo *point-and-click* ou aplicar diretamente comandos e programações, dispondo de recursos para atualização automática por meio da Web como quase nenhum outro software. Possibilita, por exemplo, que um pesquisador faça atualizações de procedimentos, comandos e códigos, utilize macros desenvolvidas por outros pesquisadores ao redor do mundo ou trabalhe com bases de dados disponíveis na internet sem que, para tanto, haja algum custo adicional.

Neste sentido, é com bastante satisfação que apresento o primeiro livro de Métodos Quantitativos Aplicados por meio do software Stata® publicado em língua portuguesa.

O livro está estruturado em nove capítulos, de acordo com o que segue:

Capítulo 1: Introdução

Capítulo 2: Estatística Descritiva, Tabelas e Gráficos

Capítulo 3: Testes de Hipótese e Análise de Variância (ANOVA)

Capítulo 4: Regressão Linear

Capítulo 5: Avaliação dos Modelos de Regressão

Capítulo 6: Regressão Robusta

Capítulo 7: Regressão Logística

Capítulo 8: Análise de Sobrevivência: Procedimento Kaplan-Meier e Regressão de Cox

Capítulo 9: Regressão com Dados em Painel

Cada capítulo está estruturado dentro de uma mesma lógica de apresentação, o que, acredito, favorece o processo de aprendizado. A aplicação de exemplos por meio da utilização do Stata® é a linha mestra, e a análise dos *outputs* gerados possibilita, em função

da teoria subjacente a cada modelagem ou técnica, um melhor entendimento do que está sendo estudado, uma vez que o passo a passo é detalhado e ilustrado e os *outputs* são analisados e interpretados sempre com caráter gerencial voltado para a tomada de decisão.

Desta maneira, acredito que o livro seja voltado tanto para pesquisadores que, por diferentes razões, se interessam especificamente por modelagem, quanto para aqueles que desejam aprofundar seus conhecimentos por meio da utilização do Stata®.

Este livro é recomendado a alunos de graduação e pós-graduação *stricto sensu* em administração, engenharia, economia, contabilidade, atuária, psicologia, medicina e saúde e demais campos do conhecimento relacionados às ciências humanas, exatas e biomédicas. É destinado também a alunos de cursos de extensão, de pós-graduação *lato sensu* e MBA´s, profissionais de empresas, consultores e demais pesquisadores que têm, como principal objetivo, o tratamento e a análise de dados estatísticos com vistas à geração de informações e ao aprimoramento do conhecimento por meio da tomada de decisão.

Aos pesquisadores que utilizarem este livro, desejo que surjam formulações de questões de pesquisa adequadas e cada vez mais interessantes, que sejam desenvolvidos modelos confiáveis, robustos e úteis à tomada de decisão, que a interpretação dos *outputs* seja mais amigável e que a utilização do Stata® resulte em importantes e valiosos frutos para novas pesquisas e novos projetos.

Aproveito para agradecer a todos que contribuíram para que este livro se tornasse realidade. Expresso aqui os mais sinceros agradecimentos aos professores da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP), da Universidade Federal do ABC (UFABC), da Fundação Instituto de Pesquisas Contábeis, Atuariais e Financeiras (FIPECAFI), da Universidade Federal de Minas Gerais (UFMG), e da Universidade Federal de São Paulo (UNIFESP), assim como aos profissionais da Montvero Consultoria e Treinamento Ltda., da StataCorp LP (College Station, Texas) e da Editora Elsevier.

Por fim, mas não menos importante, enfatizo que sempre serão muito bem-vindas contribuições, críticas e sugestões, a fim de que seja sempre possível incorporar melhorias nesta obra.

Luiz Paulo Fávero

OS AUTORES

LUIZ PAULO FÁVERO é professor livre-docente da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP) em cursos de graduação, mestrado e doutorado. É graduado em Engenharia Civil pela Escola Politécnica da USP, pós-graduado em Administração pela Fundação Getúlio Vargas (FGV/SP) e obteve os títulos de mestre e doutor em Administração pela FEA/USP. Possui Pós-Doutorado em Econometria Financeira pela Columbia University em Nova York. Participou de cursos de Gestão de Negócios pela Harvard Business School e de Técnicas de Modelagem pela California State University. É professor visitante da Universidade Federal de São Paulo (UNIFESP) e professor em cursos de pós-graduação (especialização e MBA) da FIPECAFI, da FIA e da FIPE. É membro do *Board of Directors do Global Business Research Committee*. Seus principais interesses de pesquisa situam-se na área de modelagem multivariada, econometria, otimização e estatística aplicada a finanças e economia. É autor dos livros *Análise de Dados: Modelagem Multivariada para Tomada de Decisões*, *Pesquisa Operacional para cursos de Administração, Pesquisa Operacional para cursos de Engenharia, Precificação e Comercialização Hedônica* e *Mercado Imobiliário* e coautor de *Contemporary Studies in Economics and Financial Analysis*, *Trends in International Trade Issues* e *Finanças no Varejo*. Tem publicado artigos em diversos congressos nacionais e internacionais e em periódicos científicos, incluindo *Pesquisa Operacional*, *Revista Brasileira de Estatística*, *Central European Journal of Operations Research*, *International Journal of Management*, *International Journal of Business Research*, *Global Economy and Finance Journal*, *Journal of Financial Markets and Derivatives*, *Global Business and Economics Review*, *Estudos Econômicos*, *Contabilidade e Finanças*, *RAUSP*, *Produção*, *Brazilian Business Review*, *Revista Latinoamericana de Administración*, entre outros.

PATRÍCIA BELFIORE é professora da Universidade Federal do ABC (UFABC), onde leciona disciplinas de estatística, pesquisa operacional, planejamento e controle de produção e logística para o curso de Engenharia de Gestão. É mestre em Engenharia Elétrica e doutora em Engenharia de Produção pela Escola Politécnica da Universidade de São Paulo (EPUSP). Possui Pós-Doutorado em Pesquisa Operacional e Logística pela Columbia University em Nova York. Participa de diversos projetos de pesquisa e consultoria nas áreas de modelagem, otimização e logística. Lecionou disciplinas de pesquisa operacional, análise multivariada de dados e gestão de operações e logística em cursos de graduação e mestrado no Centro Universitário da FEI e na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH/USP). Seus principais interesses de pesquisa situam-se na área de modelagem e otimização para tomada de decisões. É autora dos livros *Análise de Dados: Modelagem Multivariada para Tomada de Decisões*, *Pesquisa Operacional para cursos de Administração, Pesquisa Operacional para cursos de Engenharia* e *Redução de Custos em Logística*. Tem publicado artigos em diversos

congressos nacionais e internacionais e em periódicos científicos, incluindo *European Journal of Operational Research*, *Computers & Industrial Engineering*, *Central European Journal of Operations Research*, *International Journal of Management*, *Gestão & Produção*, *Produção*, *Transportes*, *Estudos Econômicos*, *REAd*, entre outros.

RENATA TUROLA TAKAMATSU é professora da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais (UFMG). Bacharel em Ciências Contábeis pela UFMG, mestre e doutoranda em Controladoria e Contabilidade pela Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP). Desenvolve projetos de consultoria e de pesquisa com foco em econometria e modelos multivariados aplicados. Seus principais interesses de pesquisa situam-se nas áreas de métodos quantitativos aplicados a finanças, avaliação de investimentos e teoria de carteiras, mercado financeiro e análise de demonstrações contábeis e financeiras. Tem publicado artigos em diversos congressos nacionais e internacionais e em periódicos científicos, incluindo *Modern Economy*, *Contaduría y Administración*, *Revista Universo Contábil* e *Revista Contabilidade Vista* e *Revista*, entre outros.

JANILSON ANTONIO DA SILVA SUZART é contador e professor em cursos de pós-graduação. É bacharel em Ciências Contábeis pela Universidade do Estado da Bahia (UNEB), especialista em Auditoria Pública pela UNEB, especialista em Gestão da Administração Pública e especialista em Direito da Administração Pública pela Universidade Castelo Branco/Exército Brasileiro, mestre em Contabilidade pela Universidade Federal da Bahia (UFBA) e doutorando em Controladoria e Contabilidade pela FEA/USP. Atuou como contador público em diversos órgãos federais. Seus principais interesses de pesquisa situam-se na área de contabilidade e finanças públicas, gestão e políticas públicas, contabilidade societária, sistemas de informação, análise estatística, econometria e modelagem multivariada de dados. Tem publicado artigos em diversos congressos nacionais e internacionais e em periódicos científicos, incluindo *Journal of US-China Public Administration*, *International Business Research*, *Journal of Information Systems and Technology Management*, *Administração Pública e Gestão Social*, *Contabilidade, Gestão e Governança*, *Revista Universo Contábil*, *Revista de Gestão, Finanças e Contabilidade*, *entre* outros.

CAPÍTULO 1

Introdução

1.1. VISÃO GERAL DO STATA®

O Stata® é um aplicativo estatístico que propicia a criação, a manipulação e o gerenciamento de bancos de dados, a elaboração de gráficos e as análises estatísticas. Compatível com alguns sistemas operacionais, tais como Windows®, Macintosh® ou Unix®, o programa reúne vantagens como a facilidade de utilização, as funções analíticas pré-programadas para gerenciamento dos dados e a possibilidade de programação por parte dos usuários. Essa última funcionalidade possibilita a adição de novas capacidades ao programa a partir das necessidades detectadas pelos usuários. A maioria das operações pode ser realizada via barra de comandos ou, mais diretamente, por sua digitação direta (HAMILTON, 2009).

A primeira versão do programa foi lançada em 1985 e, a partir daí, o software foi sendo desenvolvido no sentido de acompanhar as necessidades de seus usuários, angariando popularidade frente a competidores. O Stata® atualmente é utilizado por bioestatísticos, epidemiologistas, economistas, sociólogos, cientistas políticos, geógrafos, psicólogos, cientistas sociais e outros profissionais de pesquisas que se veem diante da necessidade de analisar os mais variados formatos de dados (PEVALIN; ROBSON, 2009).

O programa é capaz de utilizar fontes externas, gerar novas variáveis, combinar conjuntos de dados, sumarizá-los, além de verificar possíveis erros advindos da sua importação e/ou combinação. Além disso, é possível se trabalhar com corte transversal, longitudinal ou ambos, o que auxilia no entendimento de quaisquer aspectos inerentes ao banco de dados (BAUM, 2006).

Em termos de estatísticas, o Stata® fornece todas as ferramentas tradicionais de estatísticas univariadas, bivariadas e multivariadas, que vão desde as estatísticas descritivas e testes t até *one-way* e n-*way* ANOVA, análise de regressão e análise dos componentes principais. Além disso, o Stata® oferece um conjunto muito poderoso de técnicas de análise de variáveis dependentes qualitativas, como as técnicas de regressão probit, logit e logit multinomial. O programa oferece também funcionalidades relacionadas à análise de regressão, como a realização de testes de diagnósticos, previsão, matriz de variância e covariância robusta, além de possibilitar o uso de variáveis instrumentais e métodos como, por exemplo, o estimador dos mínimos quadrados de dois estágios (2SLS – *two-stages least squares*) e das regressões aparentemente não relacionadas (SUR – *seemingly unrelated regressions*), dentre outros (BAUM, 2006).

Estatísticas especializadas também são abrangidas de forma bastante profunda. O aplicativo inclui comandos específicos para séries temporais (ARCH – *autoregressive*

conditional heteroskedasticity, ARIMA – *autoregressive integrated moving average*, VAR – *vector autoregressive*, VEC – *vector error correction*), modelos de simulação e *bootstrapping*, estimativas de máxima verossimilhança, e mínimos quadrados não lineares. Famílias de comandos fornecem as técnicas principais utilizadas em cada uma das várias categorias: os "xt", comandos para dados em painel; e os "st", comandos para dados destinados à análise de sobrevivência.

Os gráficos do Stata® têm sido melhorados e aprimorados, possibilitando uma análise exploratória consistente dos dados e sua exportação para publicação e relatórios técnicos em diversas formas disponíveis. Cada aspecto gráfico pode ser programado e personalizado, e novos tipos de gráficos são continuamente desenvolvidos. Em adição, a capacidade de programação implica a possibilidade de geração de uma série de gráficos semelhantes, muito rapidamente (BAUM, 2006).

Usuários novos e potenciais do Stata® geralmente se questionam acerca das possíveis vantagens que esse aplicativo possui frente aos seus competidores e, principalmente, suas vantagens frente ao SPSS® (programa estatístico licenciado pela IBM® e largamente utilizado no tratamento e na análise de dados). Dentre suas vantagens, pode-se citar a aplicação de comandos mais intuitivos e com uma sintaxe mais simples. A participação de seus usuários também merece destaque, pois colaboram na criação da maior parte dos aplicativos das novas versões. Relacionado a esse ponto, tem-se o fato de que o software é conectado à internet e não há restrições de conteúdo, ou seja, é possível a instalação de novas rotinas que foram elaboradas pelos próprios usuários e que são destinadas à realização de tarefas específicas. As extensões cobrem uma vasta área de aplicação, e a possibilidade de simplesmente procurar um procedimento na internet e instalá-lo rapidamente constitui uma vantagem inegável do Stata®. Além disso, o software é particularmente amigável, quando da necessidade de análise de uma base extensa e complexa de dados (PEVALIN; ROBSON, 2009). Portanto, pode-se resumir as vantagens oferecidas pelo Stata® nos tópicos a seguir:

- Ampla utilização em pesquisas empíricas de Contabilidade, Administração, Finanças e Economia.
- Simplicidade de utilização quando comparado com ferramentas similares, como o "R" e o SAS®.
- Sintaxe simples e intuitiva.
- Possibilidade de utilização de comandos desenvolvidos por terceiros.
- Gerenciamento robusto de grandes bases de dados.

O Stata® possui menus e janelas que visam facilitar seu uso, podendo ser empregados quando se realizam procedimentos não familiares. A sintaxe do Stata® é consistente e intuitiva, o que auxilia seus usuários a trabalharem de maneira direta, tornando simples tarefas complexas e repetitivas. Os ícones e os menus, em conjunto com a janela de comandos, podem ser empregados de maneira conjunta, adaptando-se às necessidades enfrentadas pelos usuários durante a utilização do software (Figuras 1.1 e 1.2).

Figura 1.1 ***Principais janelas do Stata®, versão 12.***

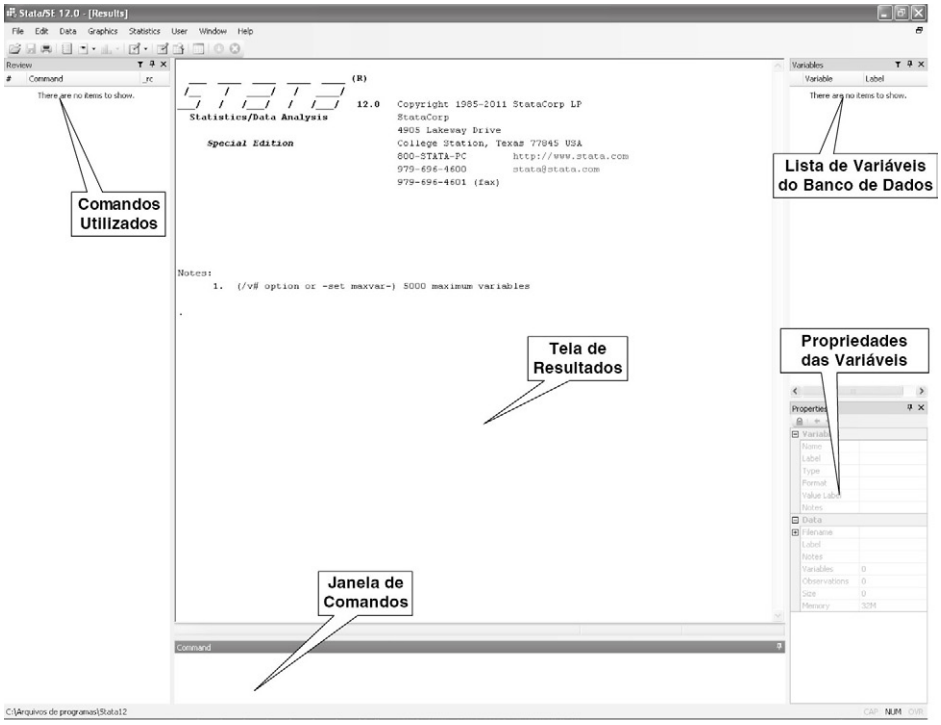


Figura 1.2 ***Componentes da tela inicial do Stata®.***

Janela de comandos

A janela de comandos (*command window*) é iniciada quando o Stata® é carregado. Por padrão, é localizada na parte inferior da tela. A janela de comandos permite que as funções sejam executadas rapidamente, mas somente se o usuário conhecer os comandos básicos.

Janela de revisão

A janela de revisão (*review window*) dos comandos utilizados é, por padrão, posicionada no canto superior esquerdo da tela. Todos os comandos são gravados nessa tela. Digitado um comando na janela de comandos, posteriormente ele será exibido e armazenado

automaticamente na janela de revisão. A janela de revisão é particularmente conveniente na análise exploratória de dados, quando o mesmo comando é utilizado com frequência para avaliar diferentes variáveis. Para reexecutar um comando, basta clicar no comando indicado na janela *Review*. Outra opção consiste na utilização da tecla PgUp (*page up*); quando o cursor estiver dentro da janela de comandos, a partir da digitação dessa tecla a sequência de comandos anteriormente executada será apresentada. O comando reaparecerá na janela de comandos, permitindo sua edição. Se um clique duplo é dado em cima do comando da janela *Review*, o Stata® irá executá-lo automaticamente.

Toda vez que executarmos uma ação via menus, automaticamente o Stata® mostrará o comando correspondente na janela de resultados. O comando use é o comando de abertura (carregamento) de arquivos.

Arquivos utilizados pelo Stata®

Os bancos de dados em Stata® possuem extensão **.dta**, sendo que existem duas versões: uma para as versões anteriores à de número 11 e outra para as versões de números 11 e 12.

Os programas (sintaxe) possuem extensão **.do** e compreendem um conjunto de comandos desenvolvidos por um usuário para automatizar a execução de determinados procedimentos. A sua visualização é possível através do uso do **do-file** *editor* (editor de *do-files*).

Os resultados (*outputs*) possuem as extensões **.log** e **.smcl**. A primeira extensão pode ser visualizada em qualquer aplicativo que manipule arquivos no formato txt. A segunda extensão, denominada log formatado para o Stata®, somente é visualizada no próprio aplicativo.

***Data Browser* e *Data Editor*: visualização e edição dos dados**

Existem diversas formas de introduzir dados no Stata®. A primeira delas consiste na digitação direta no editor de dados do Stata®. Esse editor é ativado a partir de um botão, conforme mostra a Figura 1.3. Com a ativação do editor de dados surge uma nova janela, que é uma matriz, cujas linhas representam as observações, e as colunas, as variáveis. Normalmente dados estatísticos são apresentados na forma bruta de um conjunto de indivíduos (que são as observações-linhas) com informações para diversas características (que são as variáveis-colunas).

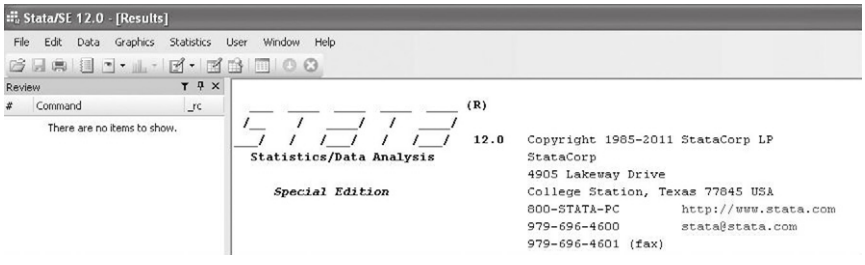


Figura 1.3 ***Tela inicial do Stata®, versão 12.***

Algumas vezes, por acidente, você pode fechar uma das janelas do Stata®. Nesse caso, basta recorrer à barra de comandos **Window** e reativar a janela. Por exemplo, caso a janela de revisão dos comandos utilizados desapareça da tela do software, é possível recuperá-la, como demonstrado na Figura 1.4.

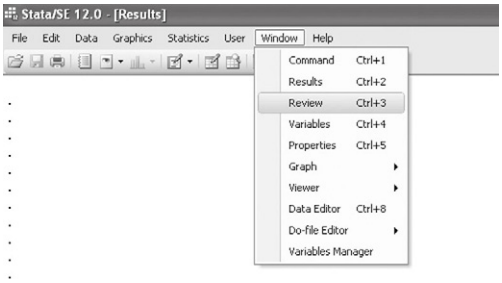


Figura 1.4 ***Acessando os comandos da barra de menus.***

Cabe destacar que o Stata® diferencia, na grafia das palavras, as letras maiúsculas e minúsculas (ou seja, é *case sensitive*). Nesse sentido, podemos citar como exemplo o comando **edit**. No Stata® o comando **edit** irá acionar a janela de edição dos dados, contudo, comandos como Edit ou EDIT não são identificados pelo programa. Seguindo nessa mesma linha de raciocínio, as variáveis *Id* e *id* seriam consideradas duas variáveis distintas.

1.2. RECURSOS NECESSÁRIOS E APLICADOS DO STATA®

1.2.1 Update

Após a instalação do software, é comum a exibição de uma caixa de texto que permite a sua atualização. Clique em **OK** e depois selecione na nova janela a opção **update all** (Figura 1.5).

Figura 1.5 ***Verificando atualizações.***

Caso essa opção não apareça, digite **update all** no *prompt* de comando (janela *command*) do Stata® (Sintaxe 1.1).

SINTAXE 1.1 Comando update.
update [query] [all]

Em que:

- query: Opção que verifica o nível de atualização da versão instalada com a versão existente no site www.stata.com.
- all: Opção que atualiza todos os comandos.

1.2.2 Background/ambiente do usuário

O plano de fundo da área de trabalho (*background*), onde estão localizadas informações sobre os comandos e os seus resultados, pode ser personalizado. Esse procedimento está disponível na opção **preferences**, disponibilizada a partir do clique inicial do botão direito do mouse na tela de resultados (Figura 1.6).

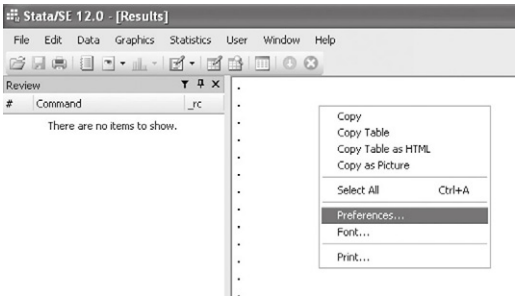


Figura 1.6 ***Acessando a opção* preferences *na tela principal.***

O Stata® oferece uma maneira de se salvar os procedimentos realizados ao longo da seção, os comandos e as tabelas de resultado. Para se iniciar a gravação do tipo log por intermédio do comando **log using** *nome_do_arquivo*, especificar o nome do arquivo no qual os comandos e resultados serão armazenados. De maneira alternativa, um arquivo **.log** pode ser criado a partir da seleção das seguintes opções na barra de menu: *File* → *Log* → *Begin*, ou ainda por intermédio do comando direto (Sintaxe 1.2).

SINTAXE 1.2 Comando log.
log [using "filename"] [close]

Em que:

- filename: Nome do arquivo no qual os resultados serão armazenados.
- close: Fechar o arquivo de log que estava sendo utilizado.

O arquivo de log pode ser criado no formato Stata (**.smcl**), ou em um formato de texto comum (**.log**). O arquivo **.smcl** (*Stata mark up and control language*) é indicado para visualização a impressão diretamente do Stata®. Esse arquivo pode controlar *hyperlinks* que auxiliem a entender os comandos ou mensagens de erro. Os arquivos do tipo **.log**, por sua vez, não exibem essa formatação, e são indicados caso se deseje inserir ou editar saídas do programa (*outputs*) em processadores de texto, tais como o Word (Figura 1.7).

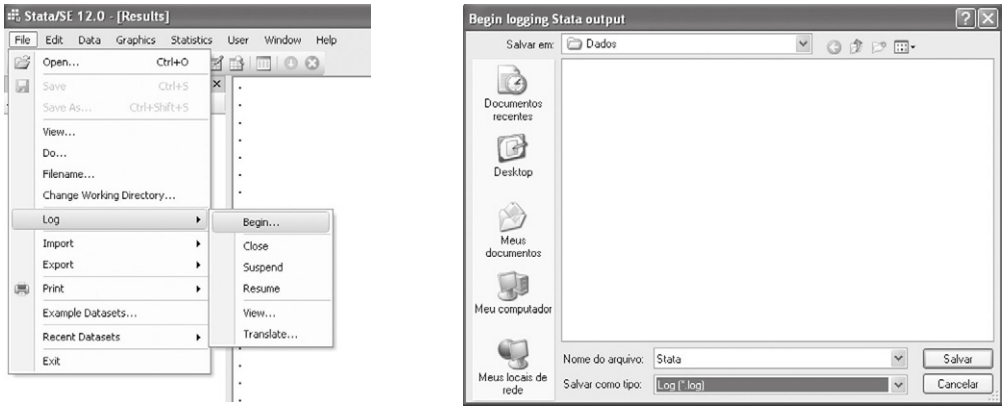


Figura 1.7 ***Gerando um arquivo de log por meio da barra de menus.***

Ao terminar de usar o Stata®, se o usuário estiver utilizando a gravação em arquivo log, é recomendável que seja fechado o respectivo arquivo com o uso do comando **log close**. Esse comando irá evitar problemas de compartilhamento do arquivo de log e garantirá que as últimas operações serão gravadas no respectivo arquivo.

O Stata® trabalha com os dados copiando-os na memória RAM. Quando o banco de dados é aberto, nenhuma mudança é realizada até que este esteja salvo. O fato de usar uma cópia dos dados é importante porque:

- Quando se utiliza o comando **use** *nome_do_arquivo*, os dados são copiados para a memória do computador, e o arquivo original é fechado (Sintaxe 1.3).

SINTAXE 1.3 Comando use.

use "filename" [, clear]

Em que:

- filename: Nome do arquivo que será aberto. Se no nome do arquivo existir algum espaço em branco é necessário utilizar aspas.
- clear: A opção **clear** somente é necessária quando já tiver sido aberta outra base de dados e desejamos simplesmente que o Stata® ignore a base aberta e passe a utilizar a base que estamos informando no comando.

- Você pode fazer o que quiser com os dados na memória, e a cópia permanente continuará a mesma em seu disco.
- A única forma de mudar uma cópia permanente dos dados é utilizando o comando **save** (Sintaxe 1.4).

SINTAXE 1.4 Comando save.
save "filename"
Em que:
- filename: Nome do arquivo que será salvo.

- Além disso, se algum erro é reportado, nenhuma mudança é realizada no banco que se encontra na memória.

1.2.3 Quantidade de memória utilizada

A definição da quantidade da memória disponível no computador a ser utilizada pelo programa constitui um aspecto importante quando da utilização de bases de dados "pesadas", que exigem muita memória. Na janela de comandos do Stata®, digite **set mem #** (Sintaxe 1.5), em que # é a quantidade de memória a ser reservada para uso das estimações durante sua sessão do Stata®.

SINTAXE 1.5 Comando set mem.
set mem #
Em que:
- #: Quantidade de memória.

Exemplo: **set mem 2m** (por exemplo, muda para 2mb a memória disponível para ser utilizada pelo aplicativo)

O Stata®, versão 12, oferece um avanço em relação às demais versões. A partir dessa versão não é mais necessário estabelecer a quantidade de memória a ser utilizada, sendo que o programa aloca a quantidade máxima de memória possível para execução dos comandos.

1.2.4 Fontes de consulta

O Stata® oferece fontes de consulta para que os usuários solucionem suas dúvidas independentemente dos níveis de dificuldade. Uma quantidade considerável de fontes sobre o aplicativo está disponível para consulta, das quais apenas a menor parcela é ligada à StatCorp (empresa responsável por criar, vender e distribuir o Stata®, além de outros produtos), sendo a maioria fornecida por uma comunidade ativa de usuários (PEVALIN; ROBSON, 2009).

Stata: <http://www.stata.com/>

No site oficial da StataCorp é possível adquirir informações sobre os produtos da StataCorp, obter suporte técnico para todas as versões do Stata®. Nos menus do Stata® é possível encontrar informações sobre encontros, treinamentos, publicações, atualizações técnicas, entre outros.

Statalist: <www.hsph.havard.edu/statalist>

O StataList é um grupo aberto de mensagens por e-mail (uma lista de discussão), sendo que qualquer interessado pode se inscrever. Existe um grande fluxo de mensagens diárias da lista, o que pode se tornar um inconveniente. Contudo, é possível escolher uma versão na qual os e-mails são condensados, reduzindo significativamente o número de mensagens recebidas. Também existem arquivos on-line do StataList que podem ser consultados.

Portal de Estatística Computacional da Universidade da Califórnia de Los Angeles (UCLA): <http://www.ats.ucla.edu/stat/stata/>

A Universidade da Califórnia possui um portal sobre o Stata®, sendo que qualquer interessado pode acessar. O site, proporcionado pela UCLA Academic Technology Service Stata Consulting Group, auxilia usuários gratuitamente. O site é uma rica fonte de notas de curso, tutoriais e exemplos detalhados que incluem comandos do Stata®, saída do programa e discussões dos *outputs* do programa.

Stata Journal: <http://www.statajournal.com>

O *Stata Journal* é um periódico publicado trimestralmente tanto em meio físico como eletrônico. Contém artigos escritos sobre o Stata®, além de adições ao software elaboradas pelos usuários, contribuindo para a evolução do programa ao longo de suas versões.

Stata Help Files

Se o usuário está interessado em um comando específico, o menu **help** o auxilia na procura de palavras-chave (*keyword*). No menu Help, é possível entender o que cada comando realiza além, de explicitar opções que podem ser combinadas. Geralmente, existem exemplos que podem auxiliar no processo de análise dos resultados (PEVALIN; ROBSON, 2009) (Sintaxe 1.6).

SINTAXE 1.6 Comando help.

help [command_or_topic_name]

Em que:

- command_or_topic_name: Comando ou assunto para o qual se deseja visualizar a ajuda do Stata®.

Por exemplo, se digitarmos, na janela de comandos, **help regression**, irá aparecer uma janela, conforme a Figura 1.8.

Figura 1.8 ***Ajuda para o tópico regression.***

O comando **findit** (Sintaxe 1.7) realiza buscas com base em determinada palavra-chave. Essas buscas envolvem tanto os arquivos de ajuda instalados no computador do

SINTAXE 1.7 Comando findit.
findit word
Em que:
- word: Termo a ser pesquisado.

usuário quanto os arquivos de ajuda on-line e das dúvidas frequentes no site do Stata®, no *Stata Journal* e nas demais fontes on-line reconhecidas pelo aplicativo. Existe também o comando **search**, apresentado na Sintaxe 1.8.

SINTAXE 1.8 Comando search.
search word
Em que:
- word: Termo a ser pesquisado.

O comando **search** é utilizado para a procura da palavra-chave na internet, enquanto o comando **net search** (Sintaxe 1.9) é utilizado para a procura por pacotes

SINTAXE 1.9 Comando net search.
net search word
Em que:
- word: Termo a ser pesquisado.

(conjunto de comandos para a realização de procedimentos específicos, como o cálculo de determinada estatística, ou para a realização de um teste) no site www.stata.com, para a instalação no computador do usuário. É possível utilizar abreviações de comandos.

Guia do Usuário do Stata® e Manual de Referência

O guia do usuário (*User's Guide*) oferece informações introdutórias do programa. O conteúdo do livro é encontrado no site, ou pode ser adquirido em conjunto com o programa. Os manuais de referência são ótimas fontes de informações estatísticas, com exemplos detalhados incluídos. Além disso, existem manuais de referência para assuntos específicos, apesar de estes variarem um pouco em função da versão utilizada do Stata®.

1.3. JANELA DE COMANDOS DO STATA®

Além da utilização de comandos, o Stata® pode ser utilizado em um modo interativo, a partir de "cliques" para aqueles que desejam utilizar o seu sistema de menus. Entretanto, mesmo ao executar os comandos por meio da barra de menus, o programa registra o comando equivalente na janela de revisão e na janela de resultados. Assim, a partir da experiência é possível aprender os comandos e posteriormente reutilizá-los ou mesmo modificá-los de maneira mais rápida.

A utilização de comandos apresenta algumas vantagens, dentre as quais a capacidade de reprodução dos resultados. Para que uma estimação possa ser considerada confiável, de maneira ideal, qualquer pessoa que acesse os mesmos programas e a mesma base de dados deverá ser capaz de reproduzir os mesmos resultados. Caso contrário, a confiabilidade da pesquisa pode ser questionada.

Em um programa de computador em que todas as ações são realizadas a partir da seleção de menus, como uma planilha, a descrição dos passos para se alcançar determinado conjunto de resultados é dificultada. A menos que cada passo e suas respectivas transformações possam ser recuperados, como garantir que os resultados com a amostra podem ser replicados em uma nova amostra? Um programa baseado em comandos possibilita a reprodução dos passos de uma estimação. Reprodutibilidade essa que facilita também a realização de análises alternativas de um modelo específico.

O Stata® possibilita a geração de um arquivo contendo apenas os comandos digitados, e o editor de **do-file** permite que a sequência de comandos ou fragmentos de programas sejam acessados, executados e salvos.

1.4. ENTRADA E MANIPULAÇÃO DE DADOS NO STATA®

O primeiro passo na análise dos dados envolve organizar os dados brutos em um arquivo no formato dos bancos de dados do Stata®.

1.4.1 Dados primários

No caso de dados primários (coletados com instrumentos próprios pelo usuário), é possível utilizar o DataEntry para criar formulários de entrada dos dados. Após a digitalização das informações, é feita a transferência dessas para um banco de dados no formato utilizado pelo Stata®. O comando utilizado será o **edit** (Sintaxe 1.10).

SINTAXE 1.10 Comando edit.
edit [varlist] [if] [in]

Em que:

- varlist: Caso não se queira editar toda a base de dados, podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

O comando **edit** pode ser acessado com o ícone "Data Editor" da barra de ferramentas.

As variáveis que aparecem na cor preta não possuem rótulos e são variáveis quantitativas. Uma variável quantitativa pode ser descrita por um número para o qual operações aritméticas, tais como média e desvio-padrão, fazem sentido. As demais variáveis (que são apresentadas em outras cores) foram consideradas variáveis qualitativas pelo programa. Variáveis qualitativas (ou categóricas), por outro lado, são simples registros de uma qualidade/característica. Dentre as variáveis qualitativas, as que aparecem na cor azul possuem o rótulo visualizado, e as que apresentam a cor vermelha são variáveis nominais (*string* ou *character*).

Uma segunda forma de se introduzir dados no Stata® é a abertura de arquivos já preparados no formato do software. Esses arquivos de dados têm uma extensão **.dta**, e utilizaremos um arquivo de exemplo que poderá ser encontrado no diretório C:\Arquivos de Programas\Stata12 denominado **auto.dta**. Para carregar esse arquivo vá até o menu *File* → *Open* e busque o arquivo **auto.dta** neste caminho.

O Stata® permite a importação ou exportação para outros formatos de bancos de dados. Por exemplo, na versão 12, é possível a importação direta de planilhas eletrônicas nos formatos utilizados pelo Excel® 97, 2003 e 2010. Em outras versões existe a possibilidade de utilização de arquivos no formato texto, no formato utilizado pelo SAS®, no formato XML (*extensible mark-up language*) ou diretamente em bases de dados relacionais (MySQL, por exemplo).

1.4.2 Stat Transfer®

Uma forma fácil de converter bancos de dados de um programa para outro é com o Stat Transfer® (Figura 1.9). Esse aplicativo pode ser considerado como um complemento aos usuários do Stata® (www.stattransfer.com) que possibilita a conversão entre diferentes formatos de dados. Dados em formatos utilizados por SPSS®, SAS® ou Excel® são convertidos para arquivos no formato reconhecido pelo Stata® facilmente. O programa possibilita a conversão de arquivos não apenas para o formato Stata®, mas entre diversos formatos de arquivo, abrangendo ampla gama de programas estatísticos e econométricos convencionalmente utilizados em Administração, Contabilidade, Economia, Engenharia, Bioestatística, entre outras áreas do conhecimento.

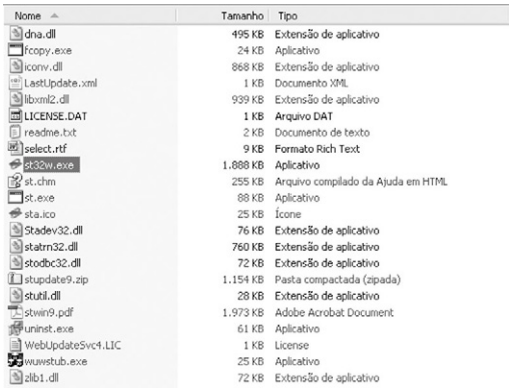


Figura 1.9 ***Acionando o Stat Transfer®.***

A Figura 1.10 mostra a tela inicial do Stat Transfer®, versão 9.

O programa apresenta duas opções de dados: o tipo de entrada de dados (***Input File Type***) e o tipo de saída dos dados (***Output File Type)***. Na primeira entrada se explicita a extensão do programa de origem dos dados, e imediatamente abaixo (***File Specification***)

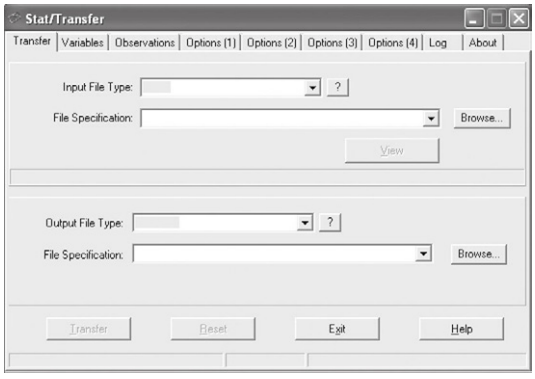


Figura 1.10 ***Tela inicial do Stat Transfer®.***

é selecionado onde o arquivo está localizado (o botão ***browse*** pode ser utilizado para a localização do arquivo).

O próximo passo é escolher a extensão do programa em que se deseja ter os dados, através da opção de saída dos dados (***Output File Type***). Uma vez selecionado o formato, na parte inferior é estabelecido onde será salvo o novo arquivo. Caso não seja alterado o local de saída dos dados, o Stat Transfer® automaticamente salvará o novo arquivo no mesmo local onde se encontram os dados originais.

Dessa maneira, é possível utilizar o Excel® para organizar bancos de dados secundários, já que esse é um programa mais acessível e com mais recursos para a edição de dados. Após a organização dos dados, o Stat Transfer® pode ser utilizado para transferir os dados para um arquivo no formato padrão do Stata®, permitindo fazer análises estatísticas mais sofisticadas.

Depois de selecionados os tipos de dados de entrada, saída, e suas respectivas localizações, é possível ativar a opção ***Transfer***, solicitando que o programa inicie a transformação dos dados para a nova extensão. Terminado o processo, é possível ver o novo arquivo criado com a extensão predefinida. Também é possível iniciar outro processo com a opção ***reset*** ou sair do programa com a opção ***Exit*** (Figura 1.11).

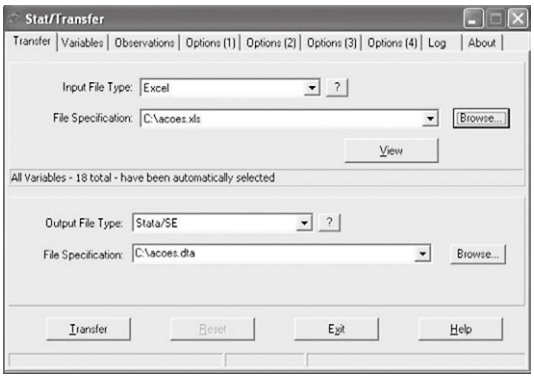


Figura 1.11 ***Janela do Stat Transfer®.***

1.4.3 Unindo duas bases de dados

Combinar dois conjuntos de dados é uma tarefa comum no gerenciamento de dados. Para realizar essa tarefa é necessário se certificar de que a estrutura de ambos os conjuntos e a lógica de organização dos dados é a mesma. O Stata® trabalha sempre com um conjunto de dados de cada vez. Porém, é possível combinar um conjunto de dados (o primeiro é denominado *master*) com outro conjunto salvo pelo usuário (denominado *using*) (Figura 1.12).

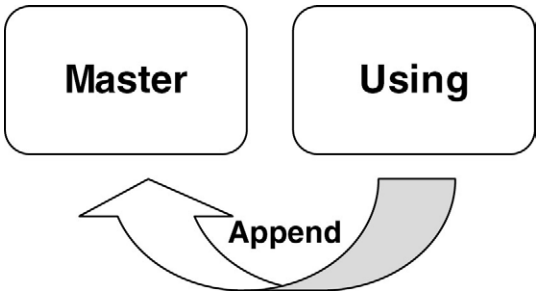


Figura 1.12 ***Unindo duas bases de dados.***

O comando **append** (Sintaxe 1.11) é utilizado para adicionar novas observações, oriundas do conjunto de dados *using*, a um conjunto de dados, denominado *master*. O comando

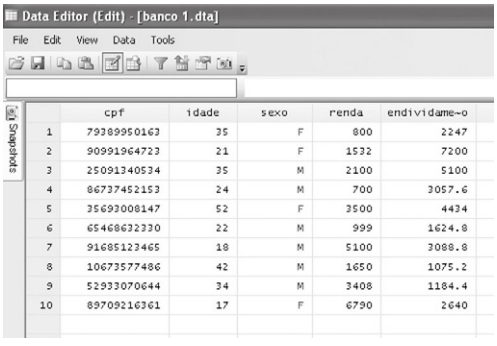
SINTAXE 1.11 Comando append.
append using filename

Em que:

- filename: Nome do arquivo que contém os dados que serão adicionados à base de dados que está aberta.

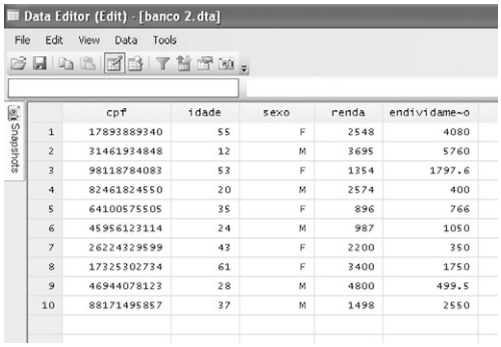
append é indicado quando as variáveis de dois bancos de dados são iguais, mas possuem observações distintas. Por exemplo, um conjunto de dados sobre pessoas de Minas Gerais pode ser adicionado ao arquivo *master* com dados sobre pessoas de São Paulo. As variáveis devem apresentar as mesmas denominações. Se uma variável aparece em apenas um dos conjuntos de dados, as demais observações serão caracterizadas como dados faltantes (*missings* ou *missing values*). A sintaxe para a execução desse tipo de procedimento é simples: basta carregar o arquivo mestre e definir para o programa qual a base de dados que será anexada.

Por exemplo, suponha que se deseje adicionar ao **arquivo banco 1** o **arquivo banco 2**. Nesse caso, o **arquivo banco 1** será considerado o arquivo *master*. Nas Figuras 1.13 e 1.14 são apresentados os dois bancos de dados.



	cpf	idade	sexo	renda	endividame~o
1	79389950163	35	F	800	2247
2	90991964723	21	F	1532	7200
3	25091340534	35	M	2100	5100
4	86737452153	24	M	700	3057.6
5	35693008147	52	F	3500	4434
6	65468632330	22	M	999	1624.8
7	91685123465	18	M	5100	3088.8
8	10673577486	42	M	1650	1075.2
9	52933070644	34	M	3408	1184.4
10	89709216361	17	F	6790	2640

Figura 1.13 ***Janela do editor de dados – arquivo* banco 1.dta.**



	cpf	idade	sexo	renda	endividame~o
1	17893889340	55	F	2548	4080
2	31461934848	12	M	3695	5760
3	98118784083	53	F	1354	1797.6
4	82461824550	20	M	2574	400
5	64100575505	35	F	896	766
6	45956123114	24	M	987	1050
7	26224329599	43	F	2200	350
8	17325302734	61	F	3400	1750
9	46944078123	28	M	4800	499.5
10	88171495857	37	M	1498	2550

Figura 1.14 ***Janela do editor de dados – arquivo* banco 2.dta.**

Aberto o arquivo mestre, basta solicitar ao programa que o arquivo desejado, no caso o **arquivo banco 2**, seja anexado, como demonstrado na Figura 1.15. Para acessar esse comando via barra de menus, clique nas seguintes opções: *Data* → *Combine datasets* → *Append datasets*.

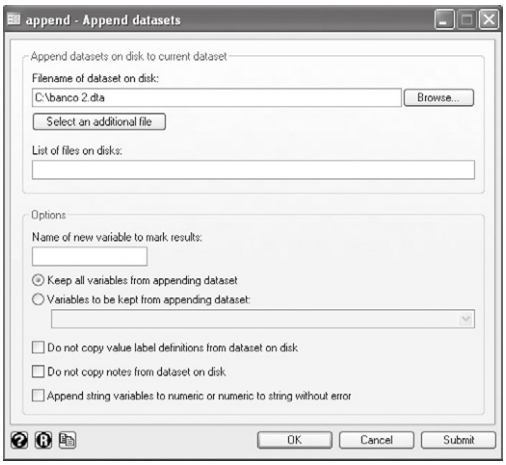
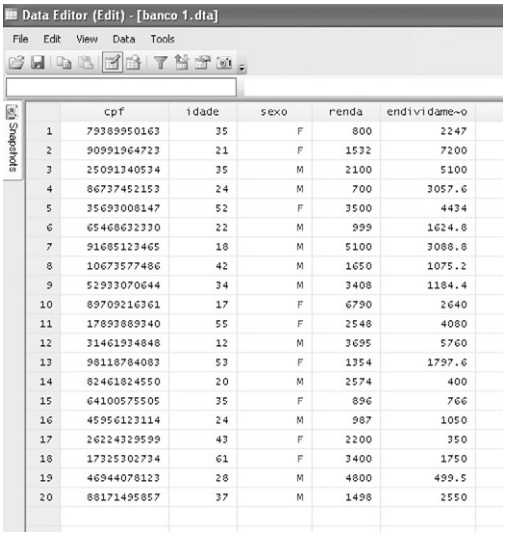


Figura 1.15 *Janela de configurações do comando* **append.**

O comando equivalente para execução desse procedimento é:

append using "C:\Documents and Settings\Meus documentos\arquivo banco 2.dta"

O resultado é um arquivo contendo 20 observações (Figura 1.16). Resultado da junção de 10 observações do arquivo 1 e 10 observações do arquivo 2.



	cpf	idade	sexo	renda	endividame~o
1	79389950163	35	F	800	2247
2	90991964723	21	F	1532	7200
3	25091340534	35	M	2100	5100
4	86737452153	24	M	700	3057.6
5	35693008147	52	F	3500	4434
6	65468632330	22	M	999	1624.8
7	91685123465	18	M	5100	3088.8
8	10673577486	42	M	1650	1075.2
9	52933070644	34	M	3408	1184.4
10	89709216361	17	F	6790	2640
11	17893889340	55	F	2548	4080
12	31461934848	12	M	3695	5760
13	98118784083	53	F	1354	1797.6
14	82461824550	20	M	2574	400
15	64100575505	35	F	896	766
16	45956123114	24	M	987	1050
17	26224329599	43	F	2200	350
18	17325302734	61	F	3400	1750
19	46944078123	28	M	4800	499.5
20	88171495857	37	M	1498	2550

Figura 1.16 *Janela do editor de dados, após o comando* **append.**

1.4.4 Mesclando duas bases de dados

O comando **merge** (Sintaxe 1.12) é responsável por fundir as observações de dois conjuntos de dados. A ideia principal desse comando é permitir a junção de dois conjuntos de dados que possuem variáveis diferentes, com exceção da variável-chave, porém, tratam

SINTAXE 1.12 Comando merge.
merge 1:1 varlist using filename
Em que:
- varlist: Lista de variáveis utilizadas como código identificador.
- filename: Nome do arquivo que contém os dados que serão adicionados à base de dados que está aberta.

da mesma observação. O comando mescla em uma mesma linha as variáveis que tenham o mesmo valor para uma variável-chave, que é utilizada como um código identificador. É muito importante, portanto, que a variável-chave tenha o mesmo formato em ambos os conjuntos de dados. Assim, por exemplo, caso se deseje fundir duas bases de dados de instituições financeiras que contenham características de clientes, pode-se ordenar essa fusão por uma variável-chave, tal como o CPF (cadastro de pessoa física) (Figura 1.17).

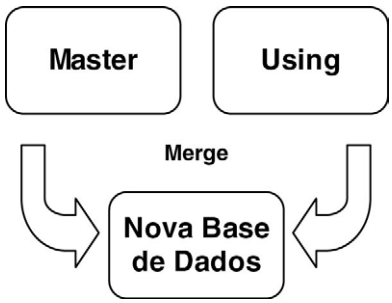


Figura 1.17 ***Mesclando duas bases de dados.***

Se as observações dos dois conjuntos de dados não coincidem, o programa apresentará campos em branco (*missing values*) para as variáveis em que a observação não encontrou correspondência. Uma vez que a viabilidade de um projeto de pesquisa depende, muitas vezes, de quantas observações realmente foi possível mesclar (por exemplo, quantas pessoas de uma base de dados de pesquisa podem ser encontradas em uma segunda base de dados), o Stata® fornece ferramentas para descobrir quantas observações realmente foram mescladas. Vamos considerar as duas bases de dados da Figura 1.18.

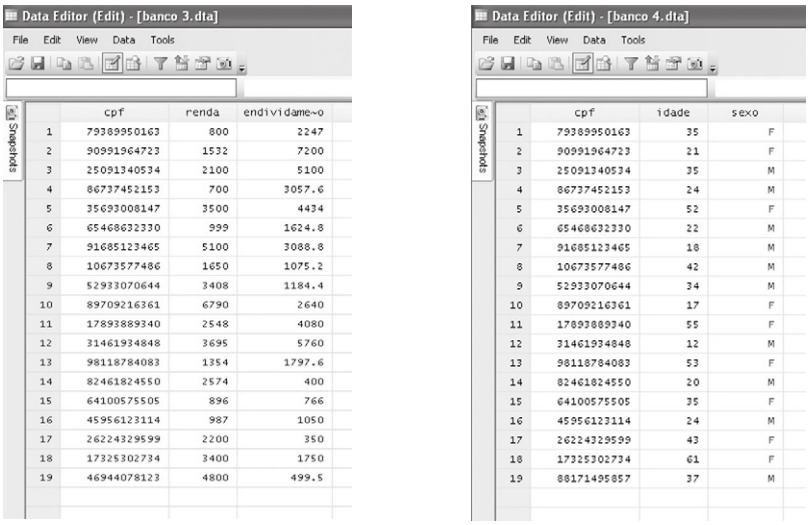


Figura 1.18 ***Janelas do editor de dados.***

O comando **merge** pode ser selecionado via barra de menus. Basta clicar nas seguintes opções: *Data* → *Combine datasets* → *Merge two datasets*. Surgirá uma janela, conforme a Figura 1.19.

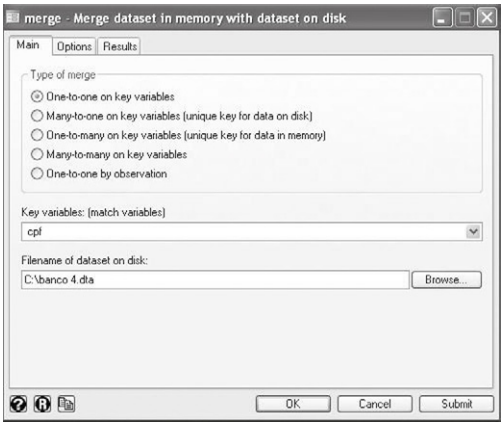


Figura 1.19 ***Janela de configurações do comando*** **merge.**

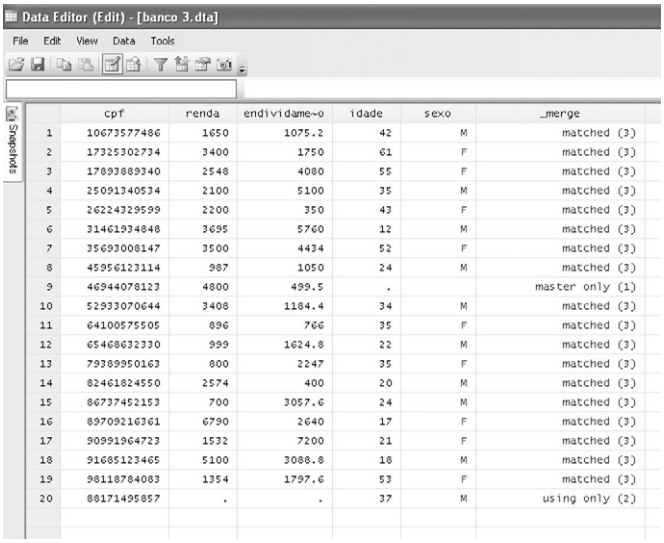
Inicialmente, será necessária a abertura do conjunto de dados que receberá os dados, o *master*. Nele serão inseridas as observações que estão no conjunto de dados *using*. No exemplo, o **arquivo banco 3** é o arquivo *master*, enquanto o **arquivo banco 4** é o conjunto de dados *using*.

A sintaxe para o comando é dada por:

merge 1:1 cpf using "C:\Documents and Settings\Meus documentos\arquivo banco 4.dta"

Esse comando irá fazer com que o Stata® adicione as informações do arquivo **banco 3.dta** ao arquivo **banco 4.dta** baseado na correspondência da variável-chave com os códigos identificadores das observações. A variável-chave não pode apresentar valores duplicados em nenhuma das bases de dados. O Stata® criará automaticamente uma nova variável denominada **_merge**.

A fusão dos dois arquivos resulta na seguinte base de dados, em que as variáveis idade e sexo (arquivo **banco 4.dta**) do segundo arquivo (arquivo **banco 3.dta**) foram fundidas com as variáveis renda e endividamento do primeiro banco de dados (Figura 1.20).



	cpf	renda	endividame~o	idade	sexo	_merge
1	10673577486	1650	1075.2	42	M	matched (3)
2	17325302734	3400	1750	61	F	matched (3)
3	17893889340	2548	4080	55	F	matched (3)
4	25091340534	2100	5100	35	M	matched (3)
5	26224329599	2200	350	43	F	matched (3)
6	31461934848	3695	5760	12	M	matched (3)
7	35693008147	3500	4434	52	F	matched (3)
8	45956123114	987	1050	24	M	matched (3)
9	46944078123	4800	499.5	.		master only (1)
10	52933070644	3408	1184.4	34	M	matched (3)
11	64100575505	896	766	35	F	matched (3)
12	65468632330	999	1624.8	22	M	matched (3)
13	79389950163	800	2247	35	F	matched (3)
14	82461824550	2574	400	20	M	matched (3)
15	86737452153	700	3057.6	24	M	matched (3)
16	89709216361	6790	2640	17	F	matched (3)
17	90991964723	1532	7200	21	F	matched (3)
18	91685123465	5100	3088.8	18	M	matched (3)
19	98118784083	1354	1797.6	53	F	matched (3)
20	88171495857	.	.	37	M	using only (2)

Figura 1.20 ***Janela do editor de dados, após o comando* merge.**

Se o valor da variável **_merge** é igual a 3 significa que existe uma correspondência entre os dois conjuntos de dados. Valores iguais a 1 ou 2 demonstram que não houve combinação entre os dois conjuntos de dados, e que a observação encontra-se apenas na primeira (*master*) ou na segunda (*using*) base de dados. Muitas vezes deseja-se manter apenas as observações que realmente foram mescladas (e onde havia informações nas duas bases de dados). Nesse caso, após a fusão dos arquivos pode-se digitar:

keep if _merge==3

O comando **keep** (Sintaxe 1.13) irá manter apenas as observações cuja variável **_merge** seja igual a 3, ou seja, onde houve correspondência entre as bases mescladas. As demais observações serão eliminadas do conjunto de dados *master*.

SINTAXE 1.13 Comando keep.
keep [varlist] [if] [in]

Em que:

- varlist: Caso não se queira utilizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

O comando **keep** pode ser acessado pela seleção das seguintes opções na barra de menus: *Data* → *Create or change data* → *Keep or drop observations*. Aparecerá uma janela, conforme a Figura 1.21.

Figura 1.21 *Janela de configurações do comando* **keep.**

1.5. VARIÁVEIS NO STATA®

Quando os dados já estão disponíveis no Stata®, alguns comandos adicionais são interessantes (Figura 1.22). O comando **drop** possibilita que variáveis e/ou observações sejam apagadas. Para exemplificar esse comando, utilizaremos o arquivo **banco 1.dta**.

Caso seja considerado que a variável renda é irrelevante na análise, pode-se excluí-la no gerenciador de variáveis (Figura 1.23).

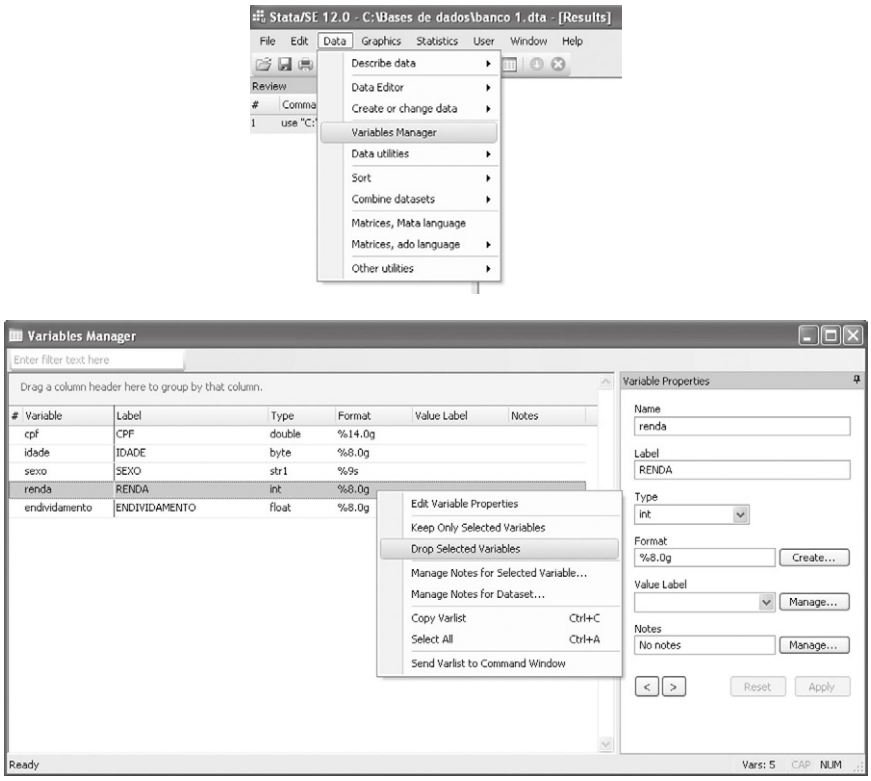


Figura 1.22 ***Acessando o gerenciador de variáveis.***

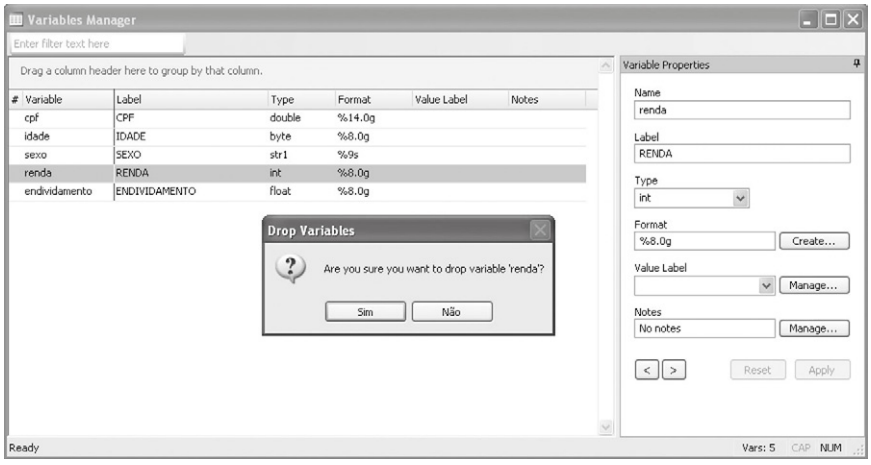


Figura 1.23 ***Confirmando no gerenciador de variáveis a exclusão de uma variável.***

O Stata® utiliza o comando **drop** (Sintaxe 1.14) para a exclusão de variáveis. Por exemplo: **drop renda**.

SINTAXE 1.14 Comando drop.
drop [varlist] [if] [in]

Em que:

- varlist: Caso não se queira utilizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Além disso, observações também podem ser excluídas pontualmente. Nesse sentido, caso se deseje remover a observação 10 por algum motivo (tal como considerá-la um *outlier*), basta solicitar a exclusão também pelo comando **drop**, da seguinte forma: **drop in 10/10**.

Via barra de menus, podemos acessar o comando **drop**, selecionando as seguintes opções: *Data* → *Create or change data* → *Keep or drop observations*. Aparecerá uma janela, conforme a Figura 1.24.

Figura 1.24 *Janela de configurações do comando* **drop.**

O comando **generate** (ou simplesmente **gen**) (Sintaxe 1.15), por sua vez, é indicado nos casos em que se deseja incluir novas variáveis, por meio de transformação de variáveis anteriormente existentes. Por exemplo, para gerar uma nova variável denominada *lnendividamento* que contém logaritmo natural do valor do endividamento, basta digitar o comando a seguir: **gen lnendividamento = log(endividamento)**.

SINTAXE 1.15 Comando generate.
generate newvar = exp [if] [in]

Em que:

- newvar: Variável que será criada.
- exp: Expressão que será utilizada na criação da variável.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Na barra de menus, esse comando está disponível em: *Data* → *Create or change data* → *Create new variable*. Surgirá uma janela, conforme a Figura 1.25.

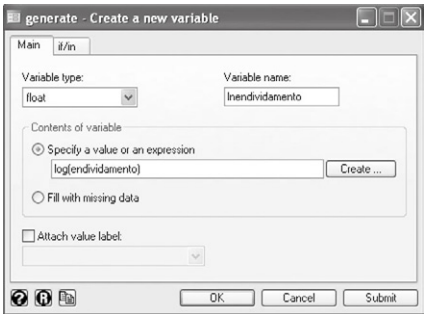


Figura 1.25 ***Janela de configurações do comando*** **generate.**

Caso queira criar uma descrição mais detalhada das variáveis, o usuário pode inserir as informações em um campo com tal destinação. A adição da descrição pode ser feita pelo comando **label var** (Sintaxe 1.16). No exemplo, deseja-se especificar na base de dados que a renda apresentada no banco de dados é a renda bruta familiar. Por exemplo: **label var renda "renda familiar bruta"**.

SINTAXE 1.16 Comando label var.
label var varname "label"

Em que:

- varname: Variável que receberá o rótulo.
- label: Rótulo atribuído à variável.

Essa opção pode ser acessada via barra de menus. Basta selecionar as seguintes opções: *Data* → *Variables Manager* (ver Figura 1.26).

Figura 1.26 *Janela de configurações do comando* **label var.**

Para visualizar uma relação das variáveis contidas na base de dados, pode ser utilizado o comando **list** (Sintaxe 1.17). Esse comando lista as variáveis, sendo que não precisam ser todas, pois o usuário pode selecionar um subgrupo. Existem diversas formas de utilização do comando **list** com o uso de "delimitadores": **if** e **in**.

SINTAXE 1.17 Comando list.

list [varlist] [if] [in]

Em que:

- varlist: Caso não se queira editar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

1.6. COMANDOS E PROGRAMAS NO STATA®

O **do-file** é uma das ferramentas mais poderosas do Stata® pela facilidade que o mesmo gera para quem utiliza o programa. No exemplo a seguir (Figura 1.27), inicialmente será aberto arquivo de dados do Stata®; pediremos para que seja: (i) computada a estatística descritiva de algumas variáveis; (ii) gerado o log de uma variável; (iii) calculada

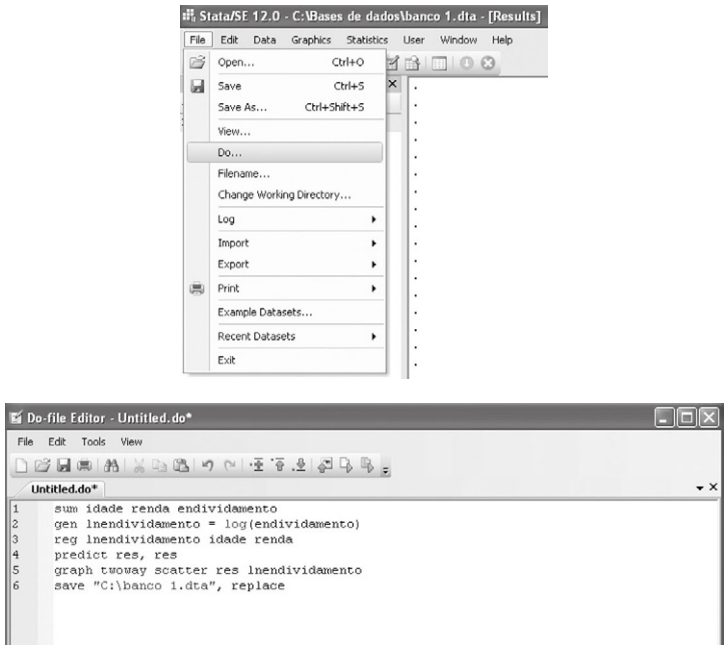


Figura 1.27 ***Acessando o*** **do-file.**

uma regressão; (iv) obtidos os resíduos do modelo e seu gráfico; e (v) salvo novamente o arquivo de dados. Todos os *do-files* podem ser salvos e armazenados, facilitando sua utilização futura.

Para se trabalhar com o **do-file**, deve-se digitar, na janela de comandos, **doedit** (Sintaxe 1.18). Os comandos a seguir devem ser digitados dentro do **do-file**. Nesse caso, basta copiar e colar para dentro da janela do **do-file**. Todos os comandos precedidos de asterisco (*) são considerados comentários.

SINTAXE 1.18 Comando doedit.

doedit [filename]

Em que:

- filename: Caso queira visualizar ou editar um arquivo de comandos, basta informar o nome do arquivo. Caso contrário, nada sendo informado o editor será aberto com um arquivo novo.

CAPÍTULO 2

Estatística Descritiva, Tabelas e Gráficos

A Estatística pode ser segregada em dois principais ramos: (i) estatística inferencial e (ii) estatística descritiva. A estatística inferencial (ou estatística indutiva) busca inferir conclusões importantes acerca da população subjacente, a partir de uma amostra representativa. Por outro lado, a estatística descritiva procura somente descrever e avaliar determinado grupo, sem tirar quaisquer conclusões ou inferências sobre um grupo maior.

Neste capítulo apresentaremos os principais comandos para a obtenção de estatísticas descritivas sobre um determinado conjunto de dados, assim como utilizaremos o Stata® para a criação de tabelas e gráficos.

Usaremos em nossos exemplos a base de dados **auto.dta**, que comumente é instalada no mesmo diretório que o Stata®. A referida base de dados possui 74 observações sobre automóveis referentes ao ano de 1978. É composta pelas variáveis contidas no Quadro 2.1.

Quadro 2.1 Variáveis que compõem a base de dados **auto.dta**

Variável	Descrição	Tipo
make	Marca e modelo	Qualitativa
price	Preço	Quantitativa
mpg	Milhagem	Quantitativa
rep78	Número de reparos no ano de 1978	Quantitativa
headroom	Potência dos alto-falantes	Quantitativa
trunk	Área do porta-malas	Quantitativa
weight	Peso	Quantitativa
length	Comprimento	Quantitativa
turn	Circunferência	Quantitativa
displacement	Deslocamento	Quantitativa
gear_ratio	Razão da engrenagem do câmbio	Quantitativa
foreign	Origem (doméstico ou estrangeiro)	Qualitativa

O primeiro passo que daremos será acionar o aplicativo Stata® e, após a sua inicialização, iremos solicitar a abertura da base de dados **auto.dta**, utilizando o comando **sysuse** (Sintaxe 2.1).

SINTAXE 2.1 Comando sysuse.

sysuse "filename" [, clear]

Em que:

- filename: Nome do arquivo que será aberto. Se no nome do arquivo existir algum espaço em branco é necessário utilizar aspas.
- clear: A opção **clear** somente é necessária quando já tiver sido aberta outra base de dados e desejamos simplesmente que o Stata® ignore a base aberta e passe a utilizar a base que estamos informando no comando.

Na janela de comandos digitaremos o seguinte:

```
sysuse auto
```

RESULTADOS 2.1 Abertura do arquivo auto.dta.

```
. sysuse auto
(1978 Automobile Data)
```

2.1. ANÁLISE EXPLORATÓRIA DE DADOS

Inicialmente buscaremos descrever os comandos que nos permitirão conhecer melhor uma base de dados. Esses comandos poderão ser utilizados para a descrição de uma base de dados por inteiro ou de algumas variáveis.

Para mostrar o sumário do banco de dados, com nome, tipo e rótulo das variáveis, vamos utilizar o comando **describe** (Sintaxe 2.2).

SINTAXE 2.2 Comando describe.

describe [varlist] [if] [in]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

No nosso exemplo, basta digitarmos o seguinte comando:
describe

RESULTADOS 2.2 Descrevendo o arquivo auto.dta.

```
. describe

Contains data from C:\Program Files\Stata12\ado\base/a/auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2011 17:45
 size:         3,182                          (_dta has notes)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
make            str18  %-18s                  Make and Model
price           int    %8.0gc                 Price
mpg             int    %8.0g                  Mileage (mpg)
rep78           int    %8.0g                  Repair Record 1978
headroom        float  %6.1f                  Headroom (in.)
trunk           int    %8.0g                  Trunk space (cu. ft.)
weight          int    %8.0gc                 Weight (lbs.)
length          int    %8.0g                  Length (in.)
turn            int    %8.0g                  Turn Circle (ft.)
displacement    int    %8.0g                  Displacement (cu. in.)
gear_ratio      float  %6.2f                  Gear Ratio
foreign         byte   %8.0g       origin     Car type
------------------------------------------------------------------------------
Sorted by:  foreign
```

Podemos, também, acionar o comando **describe** utilizando a barra de menus, basta clicarmos nas seguintes opções: *Data* → *Describe data* → *Describe data in memory*. Será exibida uma janela, conforme a Figura 2.1.

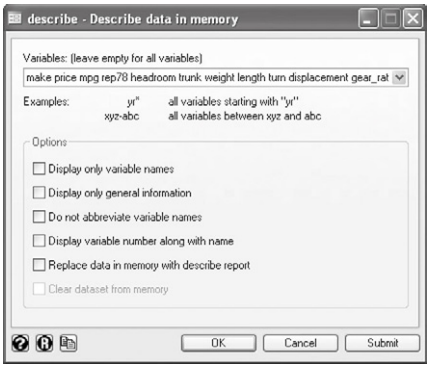


Figura 2.1 ***Janela de configurações do comando*** **describe.**

Para obtermos um resultado idêntico ao originado pelo comando que digitamos, basta deixarmos o campo *Variables* em branco e clicarmos no botão *OK*. O Stata®

possibilita que os usuários escolham algumas opções em relação ao resultado que será então fornecido.

Uma descrição mais detalhada das variáveis que compõem o banco de dados pode ser obtida por intermédio do comando **codebook** (Sintaxe 2.3).

SINTAXE 2.3 Comando codebook.

codebook [varlist] [if] [in]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Imaginemos que estamos interessados em obter mais informações sobre as variáveis *price* e *weight*. Para isso, digitaremos o seguinte comando:

```
codebook price weight
```

RESULTADOS 2.3 Obtendo informações sobre algumas variáveis.

```
. codebook price weight

-------------------------------------------------------------------------------
price                                                                   Price
-------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [3291,15906]                 units:  1
         unique values:  74                       missing .:  0/74

                  mean:   6165.26
             std. dev:    2949.5

           percentiles:        10%       25%       50%       75%       90%
                              3895      4195    5006.5      6342     11385

-------------------------------------------------------------------------------
weight                                                                  Weight
(lbs.)
-------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [1760,4840]                  units:  10
         unique values:  64                       missing .:  0/74

                  mean:   3019.46
             std. dev:    777.194

           percentiles:        10%       25%       50%       75%       90%
                              2020      2240      3190      3600      4060
```

De modo similar ao comando anterior, podemos acionar o comando **codebook** utilizando a barra de menus; basta clicarmos nas seguintes opções: *Data* → *Describe data* → *Describe data contents (codebook)*. Será exibida uma janela, conforme a Figura 2.2.

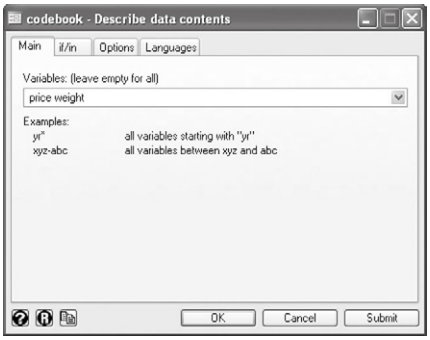


Figura 2.2 *Janela de configurações do comando* **codebook.**

Outra forma de mostrar informações sobre as variáveis da base de dados, com ilustração de quantidade de números negativos, positivos e em branco (*missing values*), além de um pequeno gráfico de ramos e folhas (com distribuição da variável entre os seus valores), é com o comando **inspect** (Sintaxe 2.4).

SINTAXE 2.4 Comando **inspect.**

inspect [varlist] [if] [in]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições de limitar a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Verificaremos agora as mesmas variáveis do exemplo anterior, *price* e *weight*. Para isso, digitaremos o seguinte comando:

inspect price weight

RESULTADOS 2.4 Inspecionando algumas variáveis.

```
. inspect price weight

price:  Price                                   Number of Observations
------------                                 ------------------------------
                                                Total    Integers   Nonintegers
|  #                              Negative          -           -           -
|  #                              Zero              -           -           -
|  #                              Positive         74          74           -
|  #                                              -----       -----       -----
|  #                              Total            74          74           -
|  #   #   .   .   .              Missing           -
+----------------------          -----
3291              15906                          74
  (74 unique values)

weight:  Weight (lbs.)                          Number of Observations
----------------------                       ------------------------------
                                                Total    Integers   Nonintegers
|  #       #                      Negative          -           -           -
|  #       #                      Zero              -           -           -
|  #   #   #   #                  Positive         74          74           -
|  #   #   #   #                                  -----       -----       -----
|  #   #   #   #                  Total            74          74           -
|  #   #   #   #   #              Missing           -
+----------------------          -----
1760              4840                           74
  (64 unique values)
```

Se desejarmos, podemos acionar o comando **inspect** utilizando a barra de menus; basta clicarmos nas seguintes opções: *Data* → *Describe data* → *Inspect variables*. Será exibida uma janela, conforme a Figura 2.3.

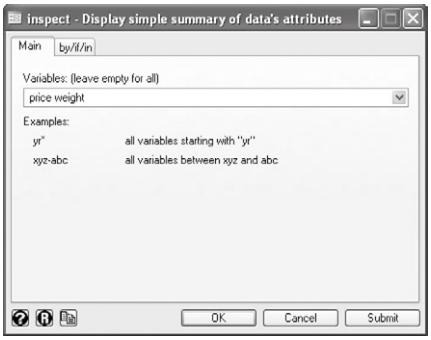


Figura 2.3 *Janela de configurações do comando* **inspect.**

Após verificamos os comandos relacionados com a obtenção de informações sobre uma base de dados ou de algumas variáveis, passaremos aos comandos que nos permitirão visualizar os dados contidos na base utilizada.

Para mostrarmos os dados da base na tela de resultados do Stata®, utilize o comando **list** (Sintaxe 2.5).

SINTAXE 2.5 Comando list.

list [varlist] [if] [in]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Para visualizar as 10 primeiras observações das variáveis *price* e *weight*, utilizaremos o seguinte comando:

list price weight in 1/10

RESULTADOS 2.5 Listando algumas observações.

```
. list price weight in 1/10

     +-----------------+
     |  price   weight |
     |-----------------|
  1. |  4,099    2,930 |
  2. |  4,749    3,350 |
  3. |  3,799    2,640 |
  4. |  4,816    3,250 |
  5. |  7,827    4,080 |
     |-----------------|
  6. |  5,788    3,670 |
  7. |  4,453    2,230 |
  8. |  5,189    3,280 |
  9. | 10,372    3,880 |
 10. |  4,082    3,400 |
     +-----------------+
```

Caso desejarmos acionar o comando **list**, por meio da barra de menus, precisaremos clicar nas seguintes opções: *Data* → *Describe data* → *List data*. Aparecerá uma janela, conforme a Figura 2.4.

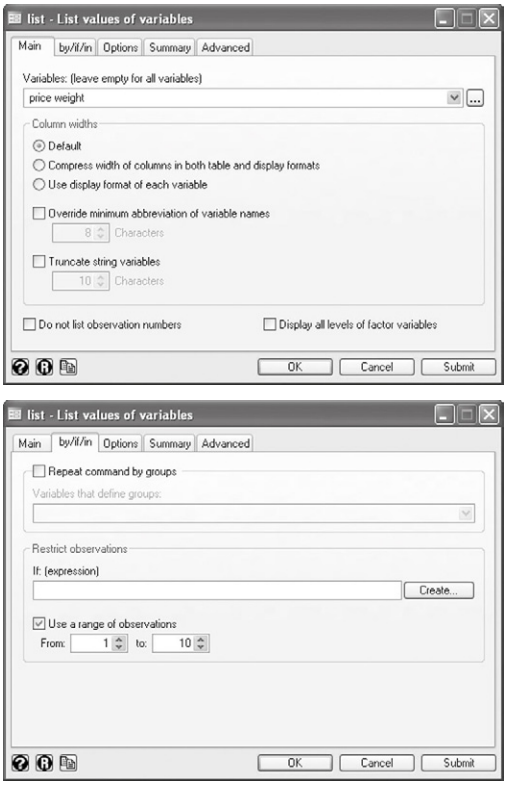


Figura 2.4 ***Janela de configurações do comando*** **list.**

Para mostrar a base de dados em uma tela separada, utilize o comando **browse** (Sintaxe 2.6).

SINTAXE 2.6 Comando browse.

browse [varlist] [if] [in]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Tendo como referência o exemplo anterior, digitaremos o seguinte comando:
browse price weight in 1/10
Na tela de resultados aparecerá o seguinte:

RESULTADOS 2.6 Exibindo algumas observações em uma janela própria.

```
. browse price weight in 1/10
```

Surgirá, então, uma janela, conforme a Figura 2.5.

Para visualizar todos os dados, podemos utilizar apenas o comando **browse**, sem opções e cláusulas. Isso também poderá ser feito utilizando a barra de menus. Basta clicarmos nas seguintes opções: *Data* → *Data Editor* → *Data Editor (Browse)*.

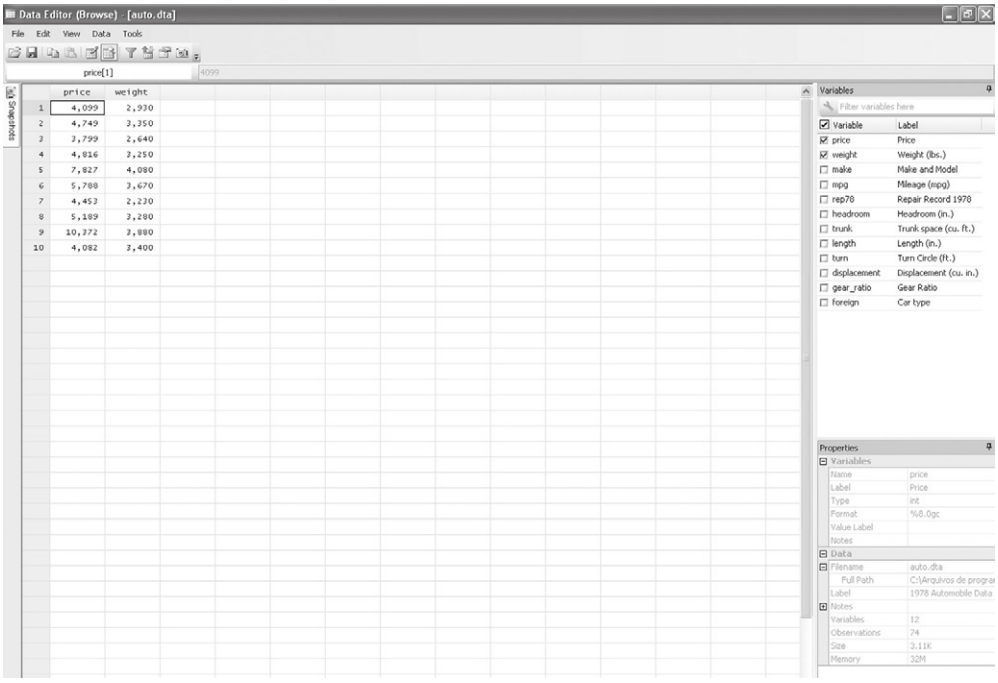


Figura 2.5 *Janela de visualização de dados – Comando* **browse.**

Caso desejássemos contar o número de observações, utilizando condições definidas com algumas das variáveis presentes na base de dados, poderíamos utilizar o comando **count** (Sintaxe 2.7).

SINTAXE 2.7 Comando count.

count [if] [in]

Em que:

- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.

Por exemplo, suponha que estamos interessados em contar apenas a quantidade de carros domésticos (na variável *foreign* o carro doméstico foi codificado com o número 0) e com preços entre 5 mil e 10 mil dólares. Dessa forma, basta especificarmos essas características no comando:

count if foreign= =0& (price> =5000 & price< =10000)

RESULTADOS 2.7 Contando observações na base de dados.

```
. count if foreign==0 & (price>=5000 & price<=10000)
  15
```

Utilizando os comandos existentes na barra de menus, podemos acionar o comando **count** da seguinte forma, bastando clicar nas seguintes opções: *Data* → *Data utilities* → *Count observations satisfying condition*. Surgirá uma janela, conforme a Figura 2.6.

Para obtermos um resultado idêntico ao originado pelo comando que digitamos, basta digitarmos as condições no campo *If*. Caso não informemos nenhuma condição, o Stata® informará o total de observações existentes na base de dados.

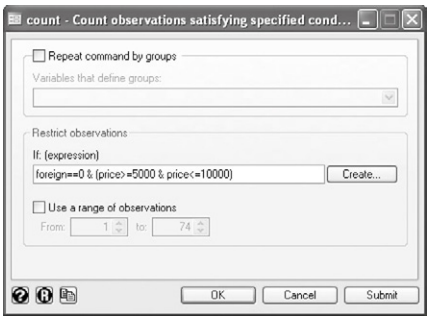


Figura 2.6 *Janela de configurações do comando* **count.**

Passaremos agora para os comandos relacionados com a obtenção de estatísticas descritivas. De um modo geral, as estatísticas descritivas estão segregadas em quatro grupos: (i) medidas de tendência; (ii) medidas de dispersão; (iii) assimetria e (iv) curtose.

O comando **summarize** (Sintaxe 2.8) apresenta estatísticas descritivas simples, tais como medianas, médias e desvios-padrão das variáveis avaliadas. Um sumário simples de estatísticas (média, desvio-padrão, valores mínimos e máximos e o número de observações) para as variáveis listadas pode ser obtido pelo comando geral.

SINTAXE 2.8 Comando **summarize.**

summarize [varlist] [if] [in] [,detail]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.
- detail: Exibe estatísticas descritivas adicionais.

Para visualizarmos um sumário com algumas estatísticas descritivas, basta digitarmos o seguinte comando:

summarize

RESULTADOS 2.8 Obtendo estatísticas descritivas da base de dados.

```
. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        make |          0
       price |         74    6165.257    2949.496      3291      15906
         mpg |         74     21.2973    5.785503        12         41
       rep78 |         69    3.405797    .9899323         1          5
    headroom |         74    2.993243    .8459948       1.5          5
-------------+---------------------------------------------------------
       trunk |         74    13.75676    4.277404         5         23
      weight |         74    3019.459    777.1936      1760       4840
      length |         74    187.9324    22.26634       142        233
        turn |         74    39.64865    4.399354        31         51
displacement |         74    197.2973    91.83722        79        425
-------------+---------------------------------------------------------
  gear_ratio |         74    3.014865    .4562871      2.19       3.89
     foreign |         74    .2972973    .4601885         0          1
```

Conforme discutido anteriormente, o Stata® irá apresentar algumas estatísticas descritivas, são elas: (i) número de observações (Obs), (ii) média (Mean), (iii) desvio-padrão (Std. Dev.), (iv) mínimo (Min) e (v) máximo (Max).

Esse comando pode ser acessado por intermédio da barra de menus. Basta que acionemos as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Summary statistics* (Figura 2.7).

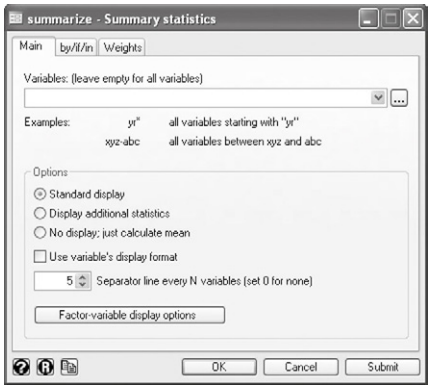


Figura 2.7 *Janela de configurações do comando* **summarize.**

Caso se deseje que na tabela sejam incluídas informações adicionais tais como percentis, variância, assimetria e curtose, a opção **detail** (precedida por uma vírgula) pode ser incluída no comando **summarize**.

O Stata® permite que alguns comandos sejam utilizados em sua forma reduzida. Em relação ao comando **summarize**, o mesmo pode ser acionado apenas digitando a sua forma reduzida **sum**.

Para visualizarmos apenas algumas variáveis (como, por exemplo *price* e *weight*) e estatísticas descritivas adicionais, utilizaremos o seguinte comando:

sum price weight, detail

RESULTADOS 2.9 Obtendo estatísticas descritivas de algumas variáveis.

```
. sum price weight, detail

                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%          3291           3291
 5%          3748           3299
10%          3895           3667       Obs                  74
25%          4195           3748       Sum of Wgt.          74

50%       5006.5                       Mean           6165.257
                        Largest       Std. Dev.      2949.496
75%          6342          13466
90%         11385          13594       Variance       8699526
95%         13466          14500       Skewness       1.653434
99%         15906          15906       Kurtosis       4.819188

                       Weight (lbs.)
-------------------------------------------------------------
      Percentiles      Smallest
 1%          1760           1760
 5%          1830           1800
10%          2020           1800       Obs                  74
25%          2240           1830       Sum of Wgt.          74

50%          3190                      Mean           3019.459
                        Largest       Std. Dev.      777.1936
75%          3600           4290
90%          4060           4330       Variance       604029.8
95%          4290           4720       Skewness       .1481164
99%          4840           4840       Kurtosis       2.118403

.
```

O Stata® irá apresentar as seguintes estatísticas descritivas: (i) número de observações (Obs), (ii) média (Mean), (iii) desvio-padrão (Std. Dev.), (iv) percentis (Percentiles), (v)

mediana (Percentiles 50%), (vi) variância (Variance), (vii) assimetria (Skewness) e (viii) curtose (Kurtosis).

Caso seja utilizada a barra de menus para se acessar o comando **summarize**, para obter as estatísticas descritivas adicionais o usuário precisará selecionar a opção '*Display additional statistics*', na janela de configuração do comando.

O Stata® permite que especifiquemos somente as estatísticas descritivas de interesse para serem exibidas na tabela. O comando para obter tal informação é o **tabstat** (Sintaxe 2.9).

SINTAXE 2.9 Comando tabstat.

tabstat varlist [if] [in] [, stats ()]

Em que:

- varlist: Caso não se queira visualizar toda a base de dados podemos informar uma lista de variáveis, separando-as por espaços em branco.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- in: A cláusula **in** (em) permite a seleção das observações de acordo com a ordem de classificação utilizada pela base de dados.
- stats: Relação de estatísticas descritivas (informadas entre parênteses) que serão exibidas no resultado.

Suponha que estamos interessados nas seguintes estatísticas descritivas da variável *price*: (i) média (mean), (ii) desvio-padrão (sd), (iii) assimetria (skewness), (iv) curtose (kurtosis), (v) número de observações (n), (vi) mínimo (min) e (vii) máximo (max). Para isso, basta informarmos na janela de comandos o seguinte:

tabstat price, stats (mean sd skewness kurtosis n min max)

RESULTADOS 2.10 Obtendo estatísticas descritivas de uma variável.

```
. tabstat price, stats (mean sd skewness kurtosis n min max)

    variable |      mean        sd  skewness  kurtosis         N       min       max
-------------+--------------------------------------------------------------------------
       price |  6165.257  2949.496  1.653434  4.819188        74      3291     15906
--------------------------------------------------------------------------------------
```

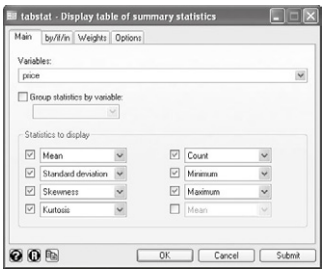


Figura 2.8 *Janela de configurações do comando* **tabstat.**

O comando **tabstat** também está acessível via barra de menus. Basta selecionarmos as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Tables* → *Table of summary statistics (tabstat)*. Aparecerá uma janela, conforme a Figura 2.8.

2.2. TESTES DE NORMALIDADE

Os testes de normalidade são bastante utilizados nos procedimentos estatísticos, muitas vezes para auxiliar o usuário na escolha do tipo de teste a ser utilizado ou para validar algum pressuposto exigido pela técnica escolhida.

Dizemos que uma variável aleatória (contínua) X apresenta distribuição normal, às vezes chamada distribuição gaussiana, quando sua função de densidade tem a seguinte forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)}{2\sigma^2}}, -\infty < x < \infty \qquad \text{[Equação 2.1]}$$

Em que μ e σ^2, conhecidos como parâmetros da distribuição, são, respectivamente, a média e a variância da distribuição. A distribuição normal é simétrica e mesocúrtica.

Existem duas formas de se testar a normalidade. A partir dos métodos gráficos é possível visualizar as distribuições de variáveis aleatórias ou as diferenças entre uma distribuição empírica e uma distribuição teórica (por exemplo, a distribuição normal padrão). Métodos numéricos apresentam estatísticas, tais como assimetria e curtose, ou realizam testes estatísticos específicos. Enquanto os métodos gráficos são intuitivos, os métodos numéricos fornecem uma maneira mais objetiva para se examinar a normalidade.

No Stata®, são necessárias utilizações de comandos individuais para obter estatísticas específicas ou esboçar gráficos. Esta seção contrasta variáveis normalmente distribuídas ou não, usando métodos gráficos e numéricos.

O histograma é o método gráfico mais amplamente utilizado. No Stata® podemos solicitar a criação de um histograma através do comando **histogram** (Sintaxe 2.10). Além disso, podemos adicionar opções, como o esboço da curva normal da variável desejada (opção **norm**).

SINTAXE 2.10 Comando histogram.

histogram varname [, norm] [, discrete]

Em que:

- varname: Nome da variável.
- norm: Caso se deseje visualizar o gráfico de densidade da função normal.
- discrete: Caso a variável não seja contínua, ou seja discreta, deve utilizar esta opção.

Vamos visualizar os histogramas das variáveis *price* e *length*. Para tanto, basta digitarmos os seguintes comandos, um de cada vez:

histogram price, norm
histogram length, norm

RESULTADOS 2.11 Gerando os histogramas das variáveis.

```
. histogram price, norm
(bin=8, start=3291, width=1576.875)

. histogram length, norm
(bin=8, start=142, width=11.375)
```

Como no Stata® os gráficos são exibidos em uma única janela, denominada *Graph*, é necessário que o usuário gere cada gráfico de uma vez e salve o gráfico gerado diretamente em um arquivo ou copiando para a memória da área de transferência.

A partir da análise gráfica, verificamos que o histograma da variável *length* está mais próximo do formato da função da distribuição normal do que o histograma da variável *price* (Figura 2.9).

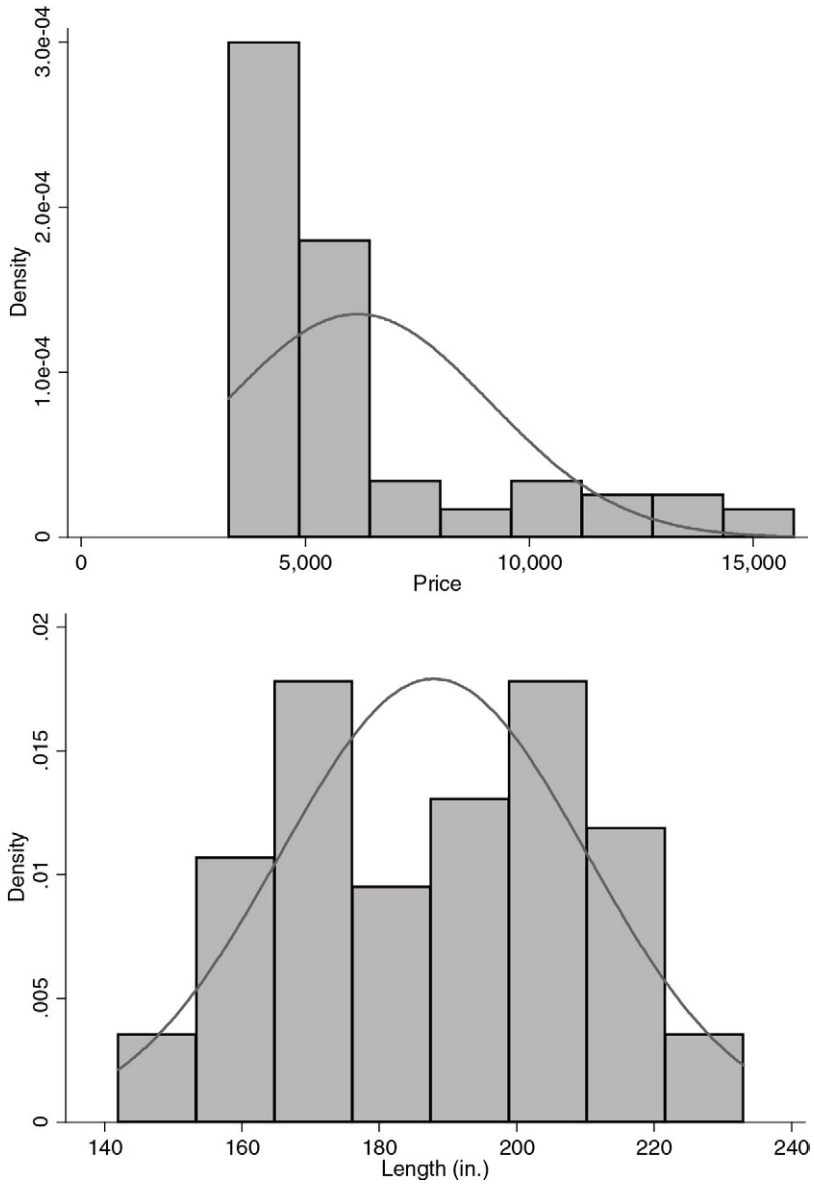


Figura 2.9 ***Histogramas das variáveis*** **price** ***e*** **length.**

Utilizando-se a barra de menus, podemos encontrar o comando **histogram**, selecionando as seguintes opções: *Graphics* → *Histogram*. Será exibida uma janela, conforme a Figura 2.10.

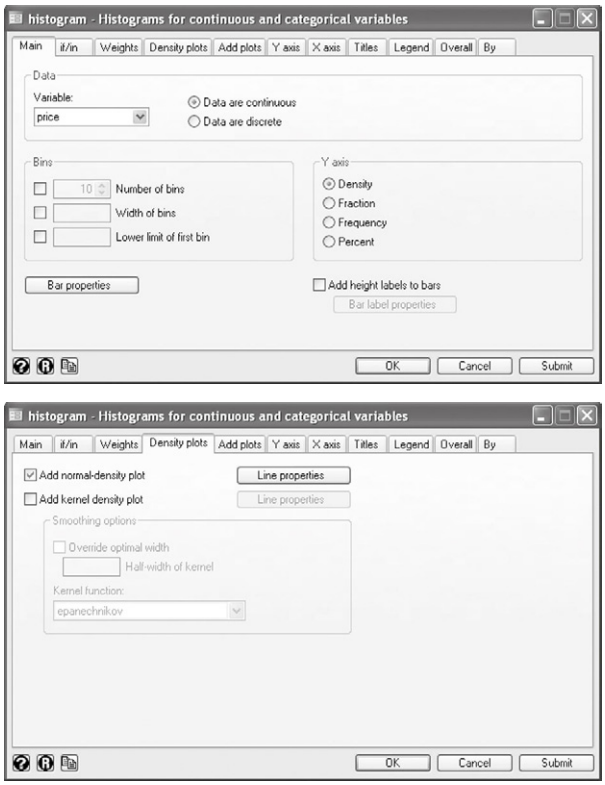


Figura 2.10 *Janela de configurações do comando* **histogram.**

O comando **graph box** (Sintaxe 2.11) esboça um box plot. Nesse gráfico, a parte sombreada representa o 25° percentil (1° quartil), a mediana (2° quartil) e o 75° percentil (3° quartil), simetricamente dispostos. O gráfico box plot pode ser utilizado para a detecção da normalidade, pois, conforme vimos anteriormente, a distribuição normal é simétrica.

SINTAXE 2.11 Comando graph box.

graph box yvars

Em que:

- yvars: Lista de variáveis, separadas por espaços em branco.

Agora, vamos visualizar os gráficos box plot para as variáveis *price* e *length* (Figura 2.11). Dessa forma, precisamos informar os seguintes comandos, um de cada vez:

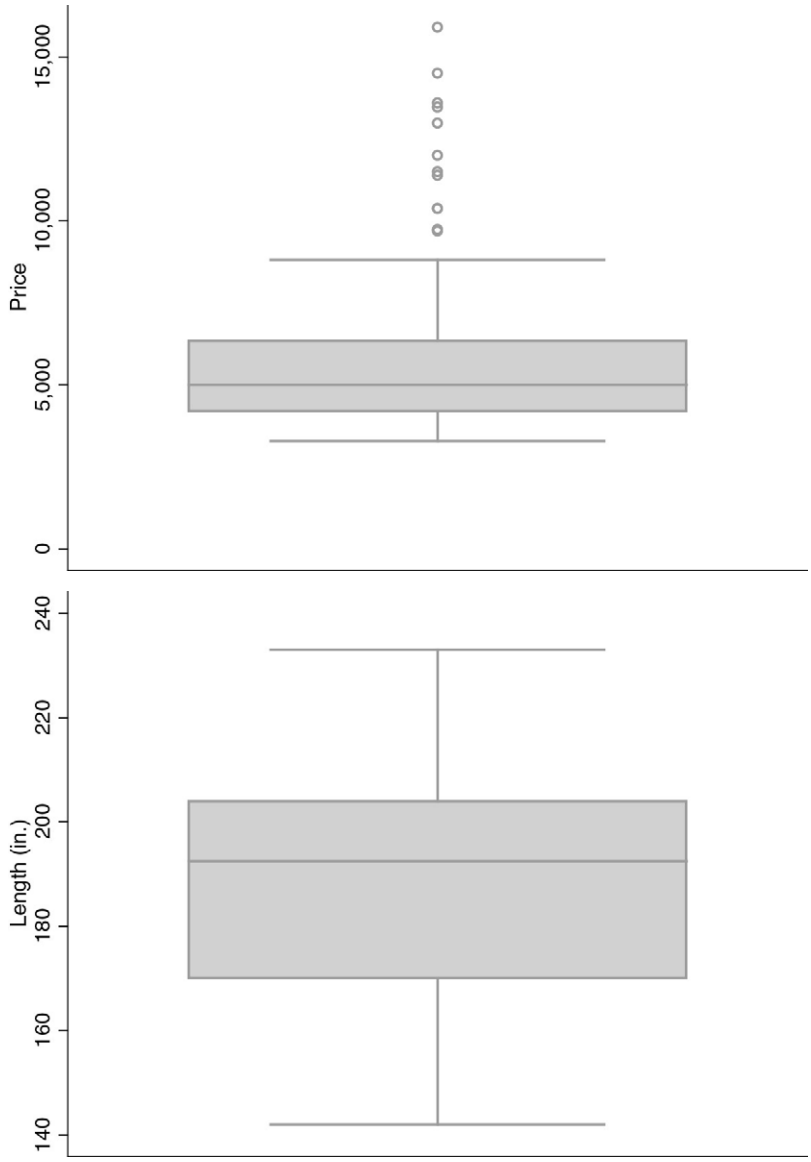


Figura 2.11 Box plot *das variáveis* price *e* length.

```
graph box price
graph box length
```

RESULTADOS 2.12 Gerando os gráficos box plot das variáveis.

```
. graph box price

. graph box length
```

A partir da análise gráfica, verificamos que o box pot da variável *length* demonstra que essa variável possui uma distribuição simétrica, enquanto a variável *price* possui uma distribuição assimétrica, pois há bastantes valores atípicos (*outliers*).

Por meio da barra de menus, podemos encontrar o comando **graph box**, selecionando as seguintes opções: *Graphics* → *Box plot*. Será exibida uma janela, conforme a Figura 2.12.

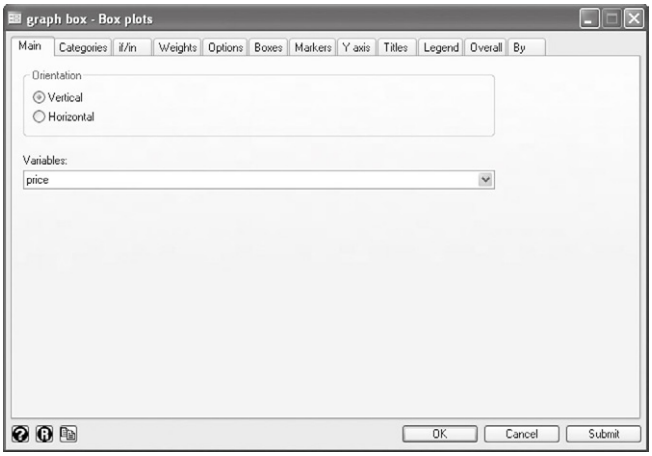


Figura 2.12 *Janela de configurações do comando* **graph box.**

A distribuição da variável em análise pode ser comparada com a função de distribuição teórica da normal. O comando **pnorm** (Sintaxe 2.12) produz um gráfico padronizado P-P plot. No Stata®, o P-P plot apresenta a distribuição cumulativa de uma variável empírica no eixo x e a distribuição teórica da normal no eixo y.

SINTAXE 2.12 Comando pnorm.

pnorm varname

Em que:

- varname: Nome da variável.

Seguindo com o nosso exemplo, vamos solicitar o gráfico P-P plot para as variáveis *price* e *length* (Figura 2.13). Novamente, lembramos que os comandos a seguir devem ser informados um de cada vez.

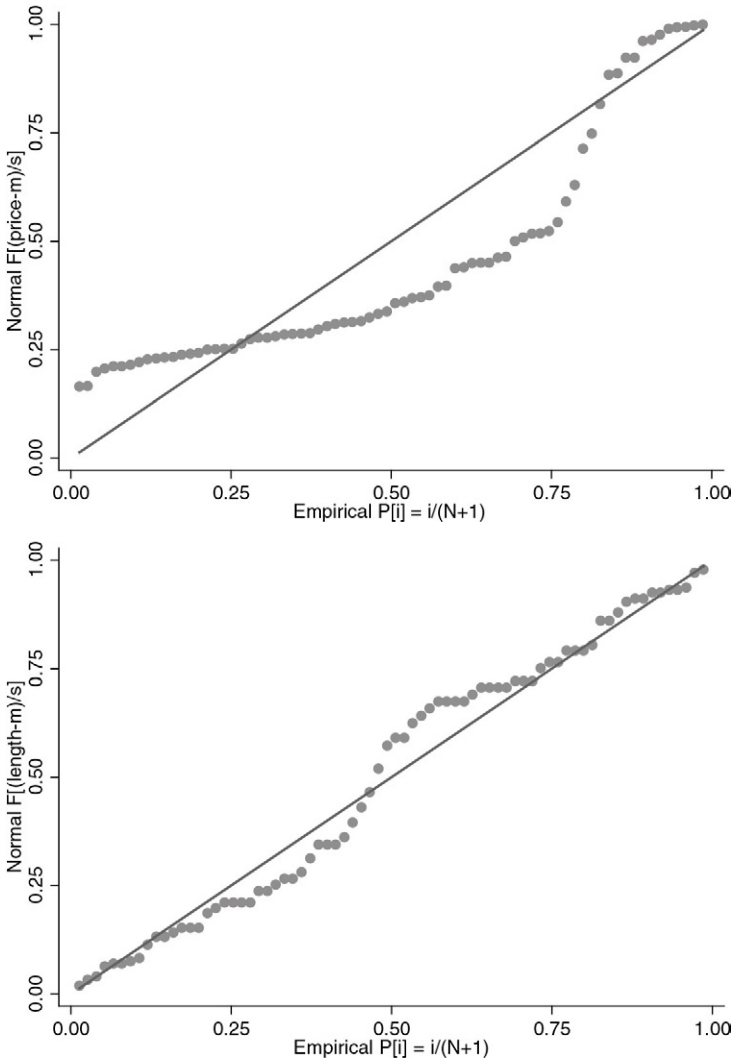


Figura 2.13 ***P-P plot das variáveis*** **price** ***e*** **length.**

```
pnorm price
pnorm length
```

RESULTADOS 2.13 Gerando os gráficos P-P plot das variáveis.

```
. pnorm price

. pnorm length
```

Analisando-se os gráficos P-P plot percebemos que o gráfico relativo à variável *price* apresenta uma forma sinuosa, desviando com muita frequência da linha estimada. Enquanto, em relação ao gráfico da variável *length*, verificamos que quase não existem desvios em comparação com a linha estimada, demonstrando, mais uma vez, que a variável estaria mais próxima de possuir uma distribuição normal.

Por intermédio da barra de menus, podemos acessar o comando **pnorm**, clicando nas seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Normal probability plot, standardized*. Na Figura 2.14 apresentamos a janela que surgirá.

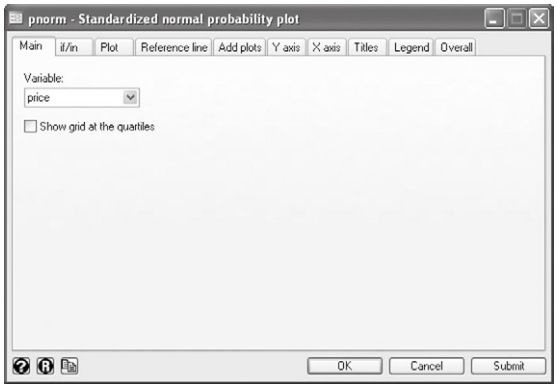


Figura 2.14 *Janela de configurações do comando* **pnorm.**

Com função similar, o gráfico Q-Q plot compara os quantis de uma distribuição de dados com os quantis da distribuição teórica da normal. O comando **qnorm** produz um gráfico Q-Q plot. O gráfico Q-Q plot apresenta um padrão similar ao gráfico P-P plot. No Stata® é acionado a partir do comando **qnorm** (Sintaxe 2.13).

SINTAXE 2.13 Comando qnorm.

qnorm varname

Em que:

- varname: Nome da variável.

Dessa vez, vamos solicitar o gráfico Q-Q plot para as variáveis *price* e *length* (Figura 2.15). Relembramos que os comandos a seguir devem ser informados um de cada vez.

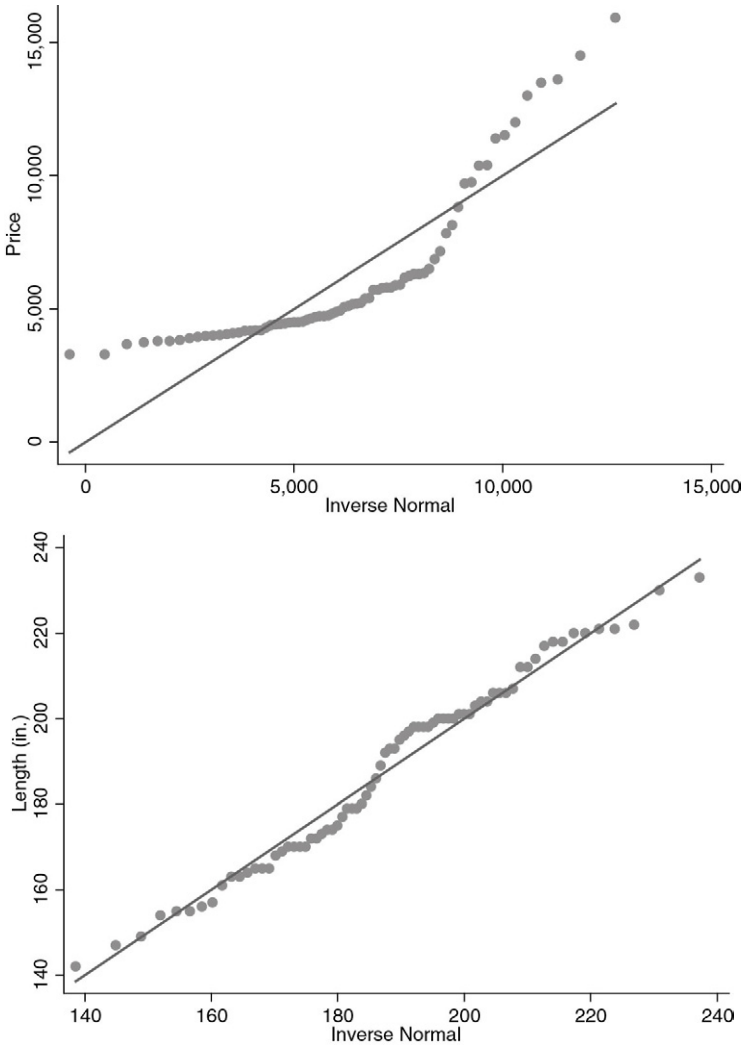


Figura 2.15 ***Q-Q plot das variáveis*** **price** ***e*** **length.**

```
qnorm price
qnorm length
```

RESULTADOS 2.14 Gerando os gráficos Q-Q plot das variáveis.

```
. qnorm price

. qnorm length
```

De maneira similar ao que ocorreu nos gráficos P-P plot, a análise dos gráficos Q-Q plot nos permite identificar que a distribuição da variável *length* é mais ajustada à distribuição teórica de uma variável normal do que a distribuição da variável *price*.

Por intermédio da barra de menus, podemos acessar o comando **qnorm**, clicando nas seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Normal quantile plot*. Na Figura 2.16 apresentamos a janela que surgirá.

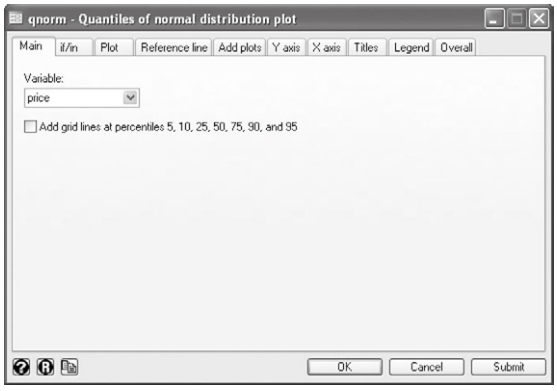


Figura 2.16 *Janela de configurações do comando* **qnorm.**

Passaremos agora aos testes estatísticos para a detecção da normalidade. Iremos descrever e demonstrar os principais testes contidos no Stata®, porém, não nos preocuparemos, nesse momento, com a análise dos resultados, pois a veremos mais detalhadamente na seção 2.5.

Para verificarmos a normalidade de uma só variável (normalidade univariada), o Stata® possui quatro métodos de teste: (i) Shapiro-Wilk, (ii) Shapiro-Francia; (iii) teste de assimetria e curtose (Skewness-Kurtosis test) e (iv) Kolmogorov-Smirnov.

Para executarmos o teste Shapiro-Wilk que, segundo Maroco (2011), é mais indicado para pequenas amostras (aquelas com até 30 observações), solicitamos o comando **swilk** (Sintaxe 2.14).

SINTAXE 2.14 Comando swilk.

swilk varlist

Em que:

- varlist: Lista de variáveis, separadas por espaços em branco.

Iremos solicitar ao Stata® que elabore o teste Shapiro-Wilk (apenas para fins didáticos, sem nos preocuparmos com a dimensão da amostra), para as variáveis *price* e *length* (Resultados 2.15). Assim, devemos digitar:

swilk price length

RESULTADOS 2.15 Teste Shapiro-Wilk.

```
. swilk price length

                    Shapiro-Wilk W test for normal data

    Variable |     Obs        W           V         z       Prob>z
-------------+--------------------------------------------------
       price |      74     0.76696      15.008     5.909     0.00000
      length |      74     0.97165       1.825     1.313     0.09461
```

O teste Shapiro-Wilk poderá ser acionado por meio da barra de menus. Para tanto, acionaremos as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Shapiro-Wilk normality test*. Surgirá a janela da Figura 2.17.

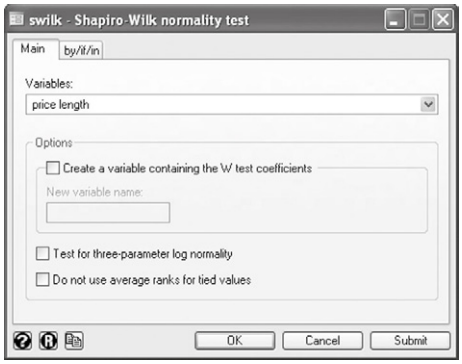


Figura 2.17 ***Janela de configurações do comando*** **swilk.**

Shapiro e Francia (1972) realizaram alterações no teste Shapiro-Wilk para que o mesmo pudesse ser utilizado com grandes amostras, dando origem ao teste Shapiro-Francia. No Stata®, esse teste é acionado pelo comando **sfrancia** (Sintaxe 2.15).

SINTAXE 2.15 Comando sfrancia.

sfrancia varlist

Em que:

- varlist: Lista de variáveis, separadas por espaços em branco.

Agora, solicitaremos que seja feito o teste Shapiro-Francia, para as variáveis *price* e *length* (Resultados 2.16).

sfrancia price length

RESULTADOS 2.16 Teste Shapiro-Francia.

```
. sfrancia price length

                  Shapiro-Francia W' test for normal data

    Variable |     Obs        W'          V'         z       Prob>z
-------------+--------------------------------------------------------
       price |      74     0.76750      16.549      5.440     0.00001
      length |      74     0.97723       1.621      0.936     0.17468
```

Assim como ocorreu com o teste Shapiro-Wilk, o teste Shapiro-Francia poderá ser acionado por meio da barra de menus. Para tanto, acionaremos as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Shapiro-Francia normality test*. Surgirá a janela da Figura 2.18.

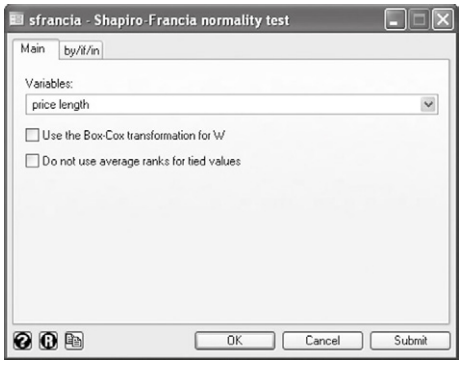


Figura 2.18 *Janela de configurações do comando* **sfrancia.**

O comando **sktest** (Sintaxe 2.16) conduz ao teste de assimetria e curtose, que é conceitualmente similar ao teste de Jarque-Bera.

SINTAXE 2.16 Comando sktest.

sktest varlist [, noadjust]

Em que:

- varlist: Lista de variáveis, separadas por espaços em branco.
- noadjust: Suprime o ajustamento empírico realizado por Royston (1991).

Executaremos o teste de assimetria e curtose, para as variáveis *price* e *length* (Resultados 2.17).

sktest price length, noadjust

RESULTADOS 2.17 Teste de assimetria e curtose.

```
. sktest price length, noadjust

                    Skewness/Kurtosis tests for Normality
                                                              ------- joint ------
    Variable |     Obs    Pr(Skewness)    Pr(Kurtosis)      chi2(2)     Prob>chi2
-------------+-------------------------------------------------------------------
       price |      74          0.0000          0.0127        28.81          0.0000
      length |      74          0.8762          0.0053         7.80          0.0202
```

Também esse comando poderá ser acionado por meio da barra de menus. Basta selecionarmos as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Skewness and kurtosis normality test*. Será exibida a janela da Figura 2.19.

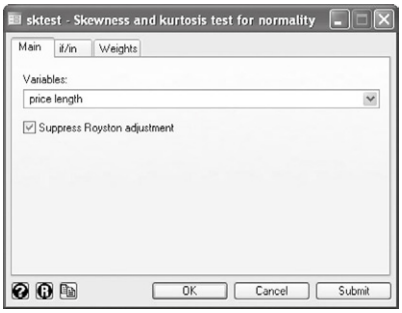


Figura 2.19 ***Janela de configurações do comando*** **sktest.**

O último teste para a detecção da normalidade univariada, disponível no Stata®, é o Kolmogorov-Smirnov. De acordo com Maroco (2011), o referido teste é indicado para grandes amostras. O teste Kolmogorov-Smirnov somente está disponível por meio do comando **ksmirnov** (Sintaxe 2.17).

SINTAXE 2.17 Comando ksmirnov.

ksmirnov varname = normal((varname-r(mean))/r(sd))

Em que:

- varname: Nome da variável.

O comando **ksmirnov**, devido à maneira como o mesmo foi construído no Stata®, exige que o comando **summarize** seja executado antes do referido comando.

Agora, iremos realizar o teste Kolmogorov-Smirnov para as variáveis *price* e *length*, utilizando os seguintes comandos:

```
summarize price
ksmirnov price = normal((price-r(mean))/r(sd))
summarize length
ksmirnov length = normal((length-r(mean))/r(sd))
```

RESULTADOS 2.18 Teste Kolmogorov-Smirnov.

```
. summarize price

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        74    6165.257    2949.496       3291      15906

.
. ksmirnov price = normal((price-r(mean))/r(sd))

One-sample Kolmogorov-Smirnov test against theoretical distribution
           normal((price-r(mean))/r(sd))

 Smaller group       D       P-value  Corrected
 ---------------------------------------------
 price:              0.2329    0.000
 Cumulative:        -0.1715    0.013
 Combined K-S:       0.2329    0.001      0.000

.
. summarize length

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      length |        74    187.9324    22.26634       142        233

.
. ksmirnov length = normal((length-r(mean))/r(sd))

One-sample Kolmogorov-Smirnov test against theoretical distribution
           normal((length-r(mean))/r(sd))

 Smaller group       D       P-value  Corrected
 ---------------------------------------------
 length:             0.0856    0.338
 Cumulative:        -0.1068    0.185
 Combined K-S:       0.1068    0.367      0.315

Note: ties exist in dataset;
      there are 47 unique values out of 74 observations.
```

O comando **ksmirnov** está disponível na barra de menus. Mesmo nessa opção o Stata® exigirá que seja executado o comando **summarize**, antes da realização do teste Kolmogorov-Sminorv. Poderá ser acessado, clicando nas seguintes opções: *Statistics* → *Nonparametric analysis* → *Tests of hypotheses* → *One-sample Kolmogorov-Smirnov test*. Aparecerá a janela da Figura 2.20.

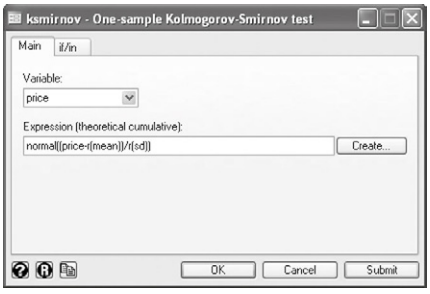


Figura 2.20 *Janela de configurações do comando* **ksmirnov.**

Técnicas de análise multivariada, tais como a análise de discriminante e a MANOVA (*multivariate analysis of variance*), exigem que as variáveis analisadas advenham de um grupo de populações que possuam uma distribuição normal multivariada. Isto significa que: (i) cada uma das variáveis é normalmente distribuída dentro do grupo, (ii) qualquer combinação linear das variáveis dependentes é normalmente distribuída, e (iii) todos os subconjuntos das variáveis devem seguir uma distribuição normal multivariada.

Um teste parcial para essa hipótese pode ser obtido com o comando **mvtest normality** (Sintaxe 2.18). O **mvtest** comando foi introduzido no Stata®, a partir da versão 11. O teste realizado é o proposto por Doornik e Hansen (2008).

SINTAXE 2.18 Comando mvtest normality.

mvtest normality varlist

Em que:

- varlist: Lista de variáveis, separadas por espaços em branco.

No próximo exemplo, iremos realizar o teste de normalidade multivariada para as variáveis *length* e *weight*, por intermédio do seguinte comando:

mvtest normality length weight

RESULTADOS 2.19 Teste Doornik-Hansen.

```
. mvtest normality length weight

Test for multivariate normality

    Doornik-Hansen                chi2(4) =    13.256    Prob>chi2 =   0.0101
```

Para acessarmos o teste Doornik-Hansen, por meio da barra de menus, devemos solicitar as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Multivariate test of means, covariances, and normality*. Surgirá a janela da Figura 2.21.

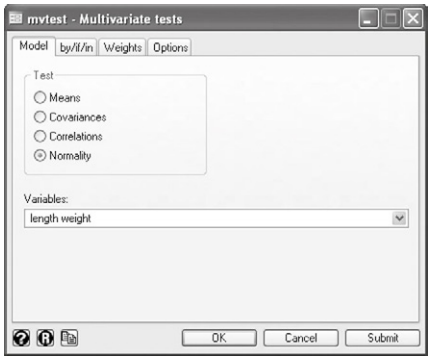


Figura 2.21 *Janela de configurações do comando* **mvtest normality.**

2.3. FREQUÊNCIA E TABULAÇÃO BIDIMENSIONAL

O comando **tabulate** (Sintaxe 2.19) é utilizado para apresentar a distribuição de frequência para os dados, excetuando-se os faltantes (*missing values*) para qualquer variável. O comando pode ser acionado por meio da sua forma reduzida **tab**.

SINTAXE 2.19 Comando tabulate para uma variável.

tabulate varname1 [, missing] [, sort] [, summarize(varname2)]

Em que:

- varname1: Nome da variável, para a qual será efetuada a tabulação.
- missing: Trata os dados faltantes como se fosse uma categoria.
- sort: Organiza a tabela de frequência, em ordem decrescente.
- summarize: Exibe estatísticas descritivas de uma variável (varname2), considerando as classes da variável que está sendo tabulada.

Primeiro, executaremos com o comando **tabulate** sem nenhuma opção para a variável *rep78*.

tabulate rep78

RESULTADOS 2.20 Tabulando em frequências uma variável.

```
. tabulate rep78

     Repair |
Record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        2.90        2.90
          2 |          8       11.59       14.49
          3 |         30       43.48       57.97
          4 |         18       26.09       84.06
          5 |         11       15.94      100.00
------------+-----------------------------------
      Total |         69      100.00
```

Para visualizarmos a quantidade de dados faltantes, iremos executar o comando **tabulate** com as opções **sort missing**.

tab rep78, sort missing

RESULTADOS 2.21 Tabulando em frequências uma variável, apresentando-se os dados faltantes.

```
. tab rep78, sort missing

     Repair |
Record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          3 |         30       40.54       40.54
          4 |         18       24.32       64.86
          5 |         11       14.86       79.73
          2 |          8       10.81       90.54
          . |          5        6.76       97.30
          1 |          2        2.70      100.00
------------+-----------------------------------
      Total |         74      100.00
```

Para acessarmos, via barra de menus, o comando **tabulate**, basta clicarmos nas seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Tables* → *One-way tables*. Aparecerá a janela da Figura 2.22.

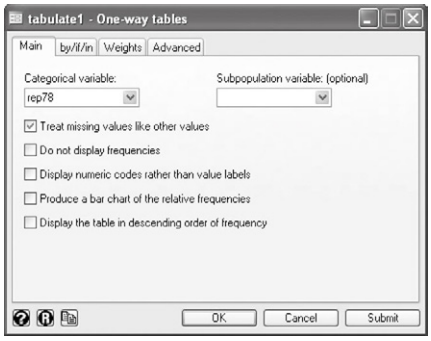


Figura 2.22 *Janela de configurações do comando* **tabulate.**

Suponha que, além de obtermos a tabela de frequências da variável *rep78,* estamos interessados em saber o comportamento da variável *price*, em cada uma das classes obtidas para a primeira variável. Para tanto, utilizaremos o seguinte comando:

tab rep78, summarize(price)

RESULTADOS 2.22 Tabulando em frequências uma variável e exibindo estatísticas descritivas de outra variável para cada classe.

```
. tab rep78, summarize(price)

     Repair |           Summary of Price
Record 1978 |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |     4,564.5   522.55191           2
          2 |   5,967.625   3,579.357           8
          3 |   6,429.233    3,525.14          30
          4 |     6,071.5   1,709.608          18
          5 |       5,913   2,615.763          11
------------+------------------------------------
      Total |   6,146.043    2,912.44          69
```

Utilizando a barra de menus, o comando **tabulate** com a opção **summarize** poderá ser acessado pelas seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Tables* → *One/two-way table of summary statistics*. Surgirá a janela da Figura 2.23.

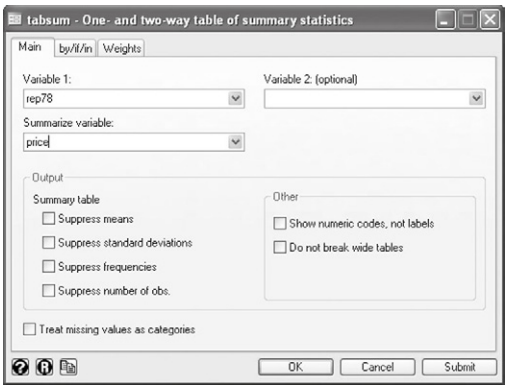


Figura 2.23 *Janela de configurações do comando* **tabulate, summarize ().**

Apesar de valiosa, a tabulação de cada variável individualmente pode não proporcionar uma riqueza de informações suficiente para se entender como duas variáveis são relacionadas. Uma tabela bivariada (*crosstab*) é simplesmente uma tabela que explicita a distribuição de uma variável ao longo das categorias de uma segunda variável. Para se criar uma tabela bivariada no Stata®, basta utilizar o comando **tabulate**, mas em vez de uma única variável, serão especificadas duas. As categorias da primeira variável estão dispostas na linha e as da segunda variável, na coluna (Sintaxe 2.20).

SINTAXE 2.20 Comando tabulate para duas variáveis.

tabulate varname1 varname2 [, missing] [, chi2] [, nofreq] [, col] [, row] [, all]

Em que:

- varname1: Nome da primeira variável.
- varname2: Nome da segunda variável.
- missing: Trata os dados faltantes como se fosse uma categoria.
- chi2: Apresenta o resultado do teste qui-quadrado de Pearson.
- nofreq: Não apresenta as frequências absolutas, apenas as relativas.
- col: Exibe as frequências apenas para a variável que está na coluna.
- row: Exibe as frequências apenas para a variável que está na linha.
- all: Apresenta todas as estatísticas disponíveis.

Vamos montar uma tabela cruzada envolvendo as variáveis *rep78* e *foreign*, utilizando o seguinte comando:

tab rep78 foreign

RESULTADOS 2.23 Tabela cruzada para duas variáveis.

```
. tab rep78 foreign

    Repair |
    Record |       Car type
      1978 |  Domestic    Foreign |     Total
-----------+----------------------+----------
         1 |         2          0 |         2
         2 |         8          0 |         8
         3 |        27          3 |        30
         4 |         9          9 |        18
         5 |         2          9 |        11
-----------+----------------------+----------
     Total |        48         21 |        69
```

Imaginemos que estamos interessados em obter somente as frequências relativas da variável *foreign* e o resultado do teste qui-quadrado de independência das variáveis *price* e *foreign*. Utilizaremos o seguinte comando:

tab rep78 foreign, chi2 nofreq col

RESULTADOS 2.24 Tabela cruzada para duas variáveis, utilizando-se opções.

```
. tab rep78 foreign, chi2 nofreq col

    Repair |
    Record |       Car type
      1978 |  Domestic    Foreign |     Total
-----------+----------------------+----------
         1 |      4.17       0.00 |      2.90
         2 |     16.67       0.00 |     11.59
         3 |     56.25      14.29 |     43.48
         4 |     18.75      42.86 |     26.09
         5 |      4.17      42.86 |     15.94
-----------+----------------------+----------
     Total |    100.00     100.00 |    100.00

          Pearson chi2(4) =  27.2640   Pr = 0.000
```

Agora, estamos interessados em produzir uma tabela cruzada que inclua dados faltantes na tabela no cálculo das porcentagens e que calcula todas as estatísticas disponíveis (qui-quadrado de Pearson, qui-quadrado da razão da verossimilhança, V de Cramer, gamma

de Kruskal e tau b de Kendall), apenas para a variável *rep78*. Para tanto, empregaremos o seguinte comando:

tab rep78 foreign, missing row all

RESULTADOS 2.25 Tabela cruzada para duas variáveis, utilizando-se opções.

```
. tab rep78 foreign, missing row all

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

    Repair |
    Record |        Car type
      1978 |  Domestic    Foreign |      Total
-----------+----------------------+----------
         1 |         2          0 |          2
           |    100.00       0.00 |     100.00
-----------+----------------------+----------
         2 |         8          0 |          8
           |    100.00       0.00 |     100.00
-----------+----------------------+----------
         3 |        27          3 |         30
           |     90.00      10.00 |     100.00
-----------+----------------------+----------
         4 |         9          9 |         18
           |     50.00      50.00 |     100.00
-----------+----------------------+----------
         5 |         2          9 |         11
           |     18.18      81.82 |     100.00
-----------+----------------------+----------
         . |         4          1 |          5
           |     80.00      20.00 |     100.00
-----------+----------------------+----------
     Total |        52         22 |         74
           |     70.27      29.73 |     100.00

          Pearson chi2(5) =  27.8735   Pr = 0.000
 likelihood-ratio chi2(5) =  30.1731   Pr = 0.000
               Cramér's V =   0.6137
                    gamma =   0.7188  ASE = 0.101
          Kendall's tau-b =   0.4540  ASE = 0.080
```

Caso se deseje acessar o comando **tabulate** para duas variáveis, podemos utilizar as seguintes opções, presentes na barra de menus: *Statistics* → *Summaries, tables, and tests* → *Tables* → *Two-way tables with measures of association*. Surgirá a janela da Figura 2.24.

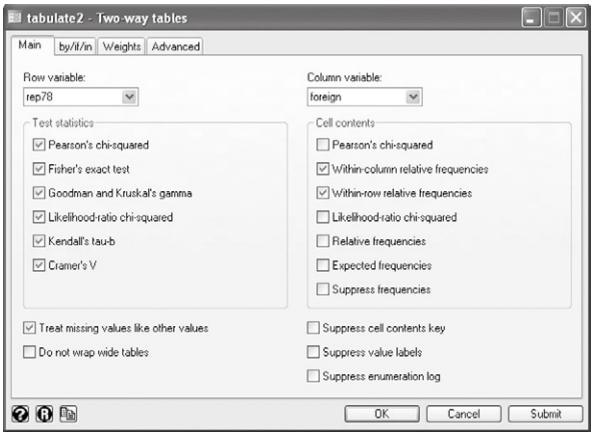


Figura 2.24 ***Janela de configurações do comando* tabulate *para duas variáveis, com opções.***

O comando **tab2** (Sintaxe 2.21) é destinado para a geração de tabelas cruzadas considerando todos os pares possíveis das variáveis informadas pelo usuário.

SINTAXE 2.21 Comando tab2.

tab2 varlist [, missing] [, chi2] [, nofreq] [, col] [, row] [, all]

Em que:

- varlist: Lista de variáveis, separadas por espaços em branco.
- missing: Trata os dados faltantes como se fosse uma categoria.
- chi2: Apresenta o resultado do teste qui-quadrado de Pearson.
- nofreq: Não apresenta as frequências absolutas, apenas as relativas.
- col: Exibe as frequências apenas para a variável que está na coluna.
- row: Exibe as frequências apenas para a variável que está na linha.
- all: Apresenta todas as estatísticas disponíveis.

Agora, vamos solicitar ao Stata® a geração de tabelas cruzadas envolvendo as variáveis *rep78*, *headroom* e *foreign*. Digitaremos o seguinte comando:

tab2 rep78 headroom foreign

RESULTADOS 2.26 Tabelas cruzadas para mais de duas variáveis.

```
. tab2 rep78 headroom foreign

-> tabulation of rep78 by headroom

    Repair |
    Record |                    Headroom (in.)
      1978 |       1.5         2.0         2.5         3.0         3.5         4.0 |     Total
-----------+------------------------------------------------------------------------+----------
         1 |         1           1           0           0           0           0 |         2
         2 |         0           3           0           0           1           2 |         8
         3 |         0           5           5           4          10           3 |        30
         4 |         2           1           5           3           2           5 |        18
         5 |         0           3           4           4           0           0 |        11
-----------+------------------------------------------------------------------------+----------
     Total |         3          13          14          11          13          10 |        69


    Repair |
    Record |     Headroom (in.)
      1978 |       4.5         5.0 |     Total
-----------+----------------------+----------
         1 |         0           0 |         2
         2 |         1           1 |         8
         3 |         3           0 |        30
         4 |         0           0 |        18
         5 |         0           0 |        11
-----------+----------------------+----------
     Total |         4           1 |        69

-> tabulation of rep78 by foreign

    Repair |
    Record |       Car type
      1978 |  Domestic    Foreign |     Total
-----------+----------------------+----------
         1 |         2           0 |         2
         2 |         8           0 |         8
         3 |        27           3 |        30
         4 |         9           9 |        18
         5 |         2           9 |        11
-----------+----------------------+----------
     Total |        48          21 |        69

-> tabulation of headroom by foreign

  Headroom |       Car type
     (in.) |  Domestic    Foreign |     Total
-----------+----------------------+----------
       1.5 |         3           1 |         4
       2.0 |        10           3 |        13
       2.5 |         4          10 |        14
       3.0 |         7           6 |        13
       3.5 |        13           2 |        15
       4.0 |        10           0 |        10
       4.5 |         4           0 |         4
       5.0 |         1           0 |         1
-----------+----------------------+----------
     Total |        52          22 |        74
```

Por meio da barra de menus, acessamos o comando **tab2**, a partir das seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Tables* → *All possible two-way tabulations*. Será exibida a janela da Figura 2.25.

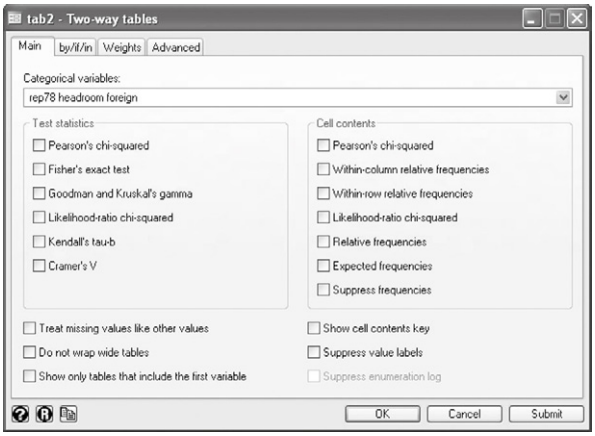


Figura 2.25 *Janela de configurações do comando* **tab2.**

2.4. OUTROS RECURSOS DA ANÁLISE EXPLORATÓRIA

O Stata® inclui um rico conjunto de ferramentas para a criação de gráficos de alta qualidade para publicação, oferecendo opções que permitem que detalhes dos gráficos sejam controlados. No entanto, em geral, os gráficos exigidos pelos usuários menos especializados, na maioria dos casos, podem ser acessados pelas configurações-padrão do Stata®.

Além disso, a interface gráfica do Stata® organiza as opções de gráficos diferentes de uma forma intuitiva, proporcionando seu acesso sem que a sintaxe de cada opção seja memorizada. Isso não significa que não é interessante salvar os comandos, mas, sim, que, para gráficos complexos, a interface gráfica auxilia a identificação de tais comandos.

O Stata® também possui um editor de gráficos que possibilita sua modificação mesmo depois que o gráfico tenha sido criado. Isto oferece um maior controle, mesmo que nessa edição não seja exibido o comando equivalente às modificações para que o gráfico seja executado novamente.

Se, posteriormente, forem necessárias quaisquer alterações nos dados, será necessário que o gráfico seja criado novamente. Dessa maneira, o gráfico, sempre que possível, deve ser criado com todas as configurações desejadas. Mesmo assim, o editor ainda pode ser considerado uma ferramenta muito útil.

A criação de gráficos não altera os dados armazenados, logo, o pior que pode acontecer é o gráfico ser mal esboçado, o que o torna inutilizável.

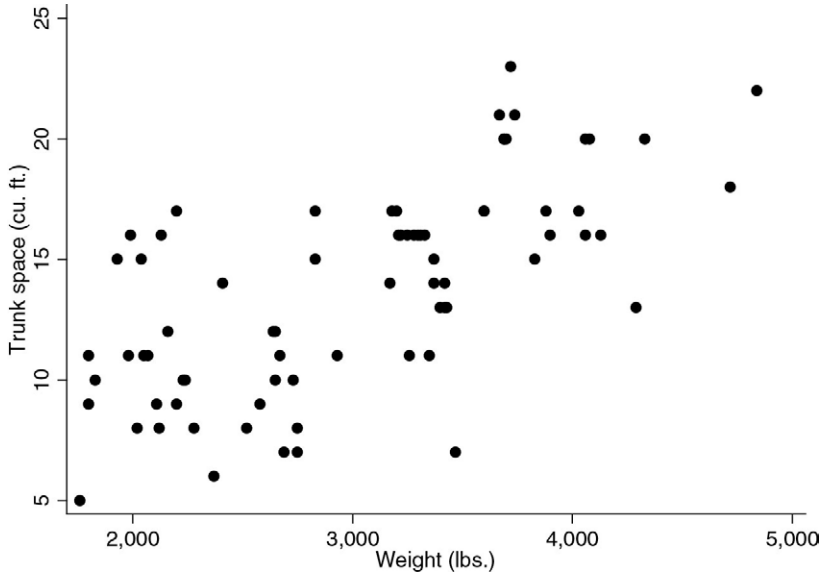


Figura 2.26 ***Gráfico de dispersão entre as variáveis*** **trunk** ***e*** **weight.**

Vamos começar com um gráfico de dispersão simples, em que a área do porta-malas (*trunk*) é definida como a variável Y e o peso (*weight*), como variável X (Figura 2.26). O Stata® refere-se a qualquer gráfico em que existem as variáveis Y e X como um gráfico **twoway** (Sintaxe 2.22).

SINTAXE 2.22 Comando **twoway**.

twoway plot varname1 varname2 [if] [, by(varname3)] [, sort]

Em que:

- plot: Tipo de gráfico que será gerado (scatter, line, bar, lfit, qfit, lfitci e qfitci são alguns dos gráficos disponíveis).
- varname1: Nome da primeira variável, que ficará no eixo Y.
- varname2: Nome da segunda variável, que ficará no eixo X.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- by: A opção **by** permite a geração de gráficos independentes para cada subpopulação, em um mesmo gráfico, considerando a variável varname3.
- sort: Organiza os dados das variáveis, em ordem crescente.

Na janela de comandos do Stata®, digitaremos o seguinte comando:
twoway scatter trunk weight

RESULTADOS 2.27 Gerando gráfico de dispersão.

```
. twoway scatter trunk weight
```

Caso desejássemos adicionar uma segunda variável no eixo Y no diagrama de dispersão, como por exemplo a variável *mpg* (Figura 2.27), basta adicionarmos um novo gráfico entre parênteses ao comando, com a mesma variável X (*weight*) mas com uma diferente variável Y. Outra opção é separar os comandos com o símbolo **||**. Assim, digitaremos no Stata® o seguinte comando:

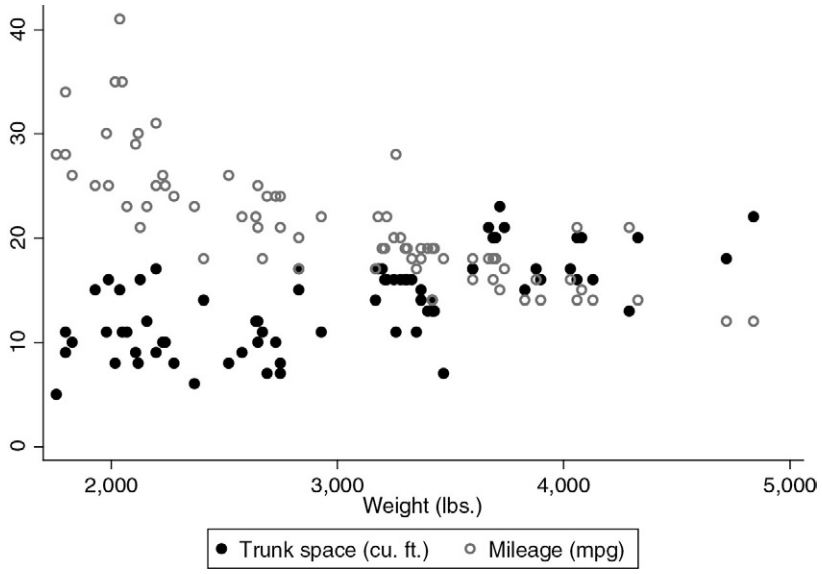


Figura 2.27 ***Gráfico de dispersão entre as variáveis*** **trunk, mpg *e* weight.**

twoway (scatter trunk weight) (scatter mpg weight)
ou
twoway scatter trunk weight || scatter mpg weight

RESULTADOS 2.28 Gerando gráfico de dispersão para dois pares de variáveis.

```
. twoway (scatter trunk weight) (scatter mpg weight)
```

Podemos desejar incluir apenas um grupo específico de observações, que pode ser especificado pelo comando **if** (Resultados 2.29 e Figura 2.28). No nosso exemplo, essa opção pode ser especificada conforme o seguinte comando, caso se deseje apenas plotar carros nacionais.

twoway (scatter trunk weight) (scatter mpg weight) if foreign = =0

RESULTADOS 2.29 Gerando gráfico de dispersão para dois pares de variáveis, com o uso da opção if.

```
. twoway (scatter trunk weight) (scatter mpg weight) if foreign==0
```

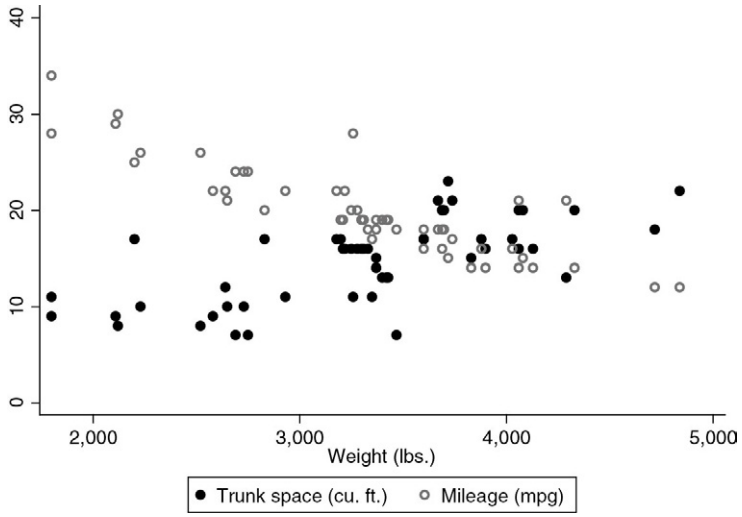


Figura 2.28 ***Gráfico de dispersão entre as variáveis*** **trunk, mpg** ***e*** **weight*****, utilizando-se a opção*** **if.**

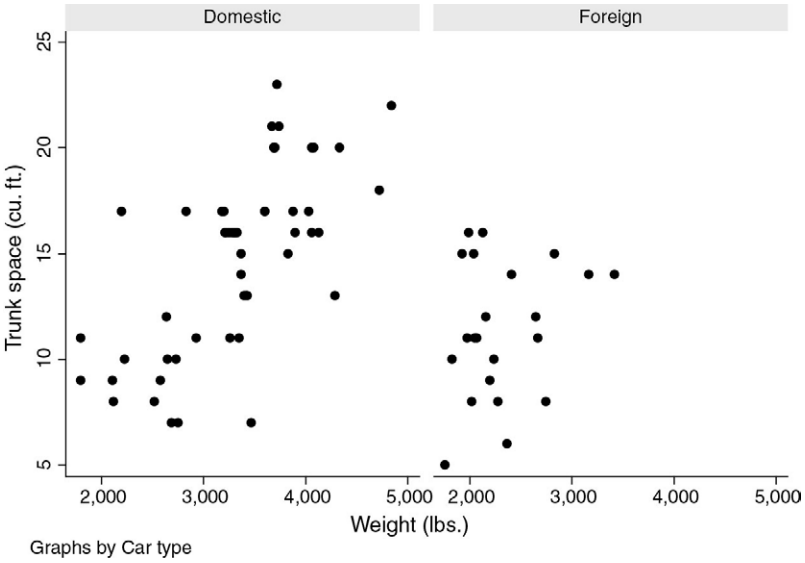


Figura 2.29 ***Gráfico de dispersão entre as variáveis*** **trunk** ***e*** **weight*****, utilizando-se a opção*** **by.**

Utilizando a opção **by** (Figura 2.29) no comando **twoway**, é esboçada separadamente cada subpopulação em um mesmo gráfico. Nesse sentido, por exemplo, para obtermos separadamente a relação entre a área do porta-malas e o peso do veículo, especificamente por nacionalidade, o comando é:

twoway scatter trunk weight, by(foreign)

RESULTADOS 2.30 Gerando gráfico de dispersão para duas variáveis, com o uso da opção by.

```
. twoway scatter trunk weight, by(foreign)
```

Voltando ao gráfico no qual se explicita a relação entre o tamanho do porta-malas e o peso do veículo, podemos desejar conectar os pontos. Nesse caso, em vez de se solicitar um gráfico de dispersão (**scatter**), podemos solicitar um gráfico de linha (**line**) (Resultados 2.31 e Figura 2.30), por meio do seguinte comando:

twoway line trunk weight

Provavelmente, o gráfico não se apresentou como o esperado: de fato, o gráfico aparenta ser somente um monte de rabiscos. Isso porque, por padrão, o Stata®

RESULTADOS 2.31 Gerando gráfico de linha para duas variáveis.

```
. twoway line trunk weight
```

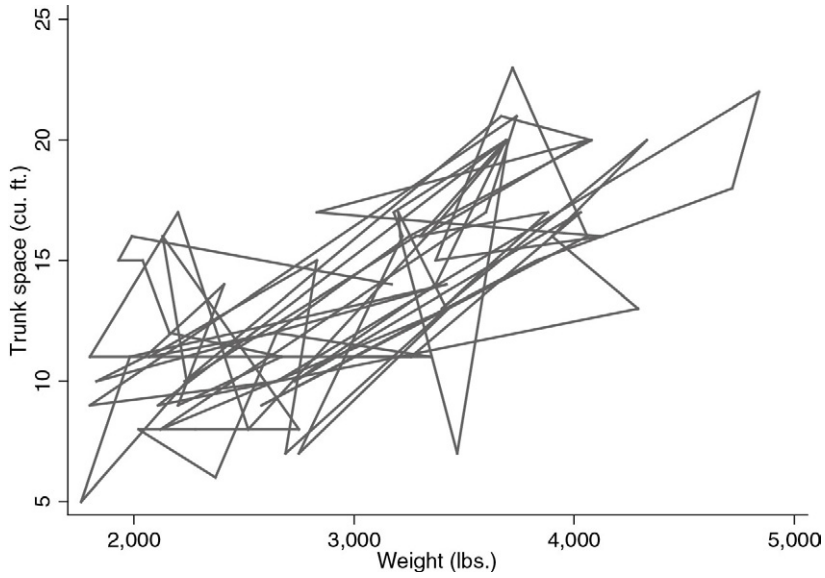


Figura 2.30 ***Gráfico de linha entre as variáveis*** **trunk** ***e*** **weight.**

estabelece a ligação entre a observação um para a dois, e da observação dois para a três, e assim por diante, seguindo a ordem no banco de dados. Contudo, o que realmente desejamos é que sejam ligados o veículo com menor peso com o próximo de menor peso. Portanto, deve-se explicitar essa opção por intermédio da opção **sort** (Resultados 2.32 e Figura 2.31).

twoway line trunk weight, sort

RESULTADOS 2.32 Gerando gráfico de linha para duas variáveis, com a opção sort.

```
. twoway line trunk weight, sort
```

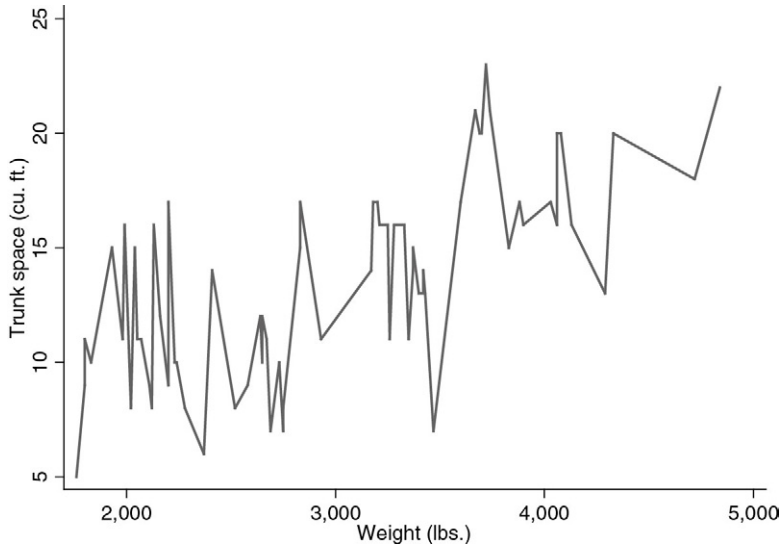


Figura 2.31 ***Gráfico de linha entre as variáveis*** **trunk** ***e*** **weight*****, utilizando a opção*** **sort.**

O Stata® pode plotar diferentes tipos de linha de ajustamento automaticamente. As mais comuns estão associadas aos comandos **lfit** (tendência linear), **qfit** (tendência quadrática), **lfitci** (tendência linear com intervalos de confiança) e **qfitci** (tendência quadrática com intervalos de confiança). Eles não são muito interessantes por si sós, mas geralmente são sobrepostos a um gráfico de dispersão.

Por exemplo, suponha que queiramos visualizar a reta linear que relaciona a variável *mpg* com a variável *weight* (Figura 2.32). Para isso, utilizaremos o seguinte comando:

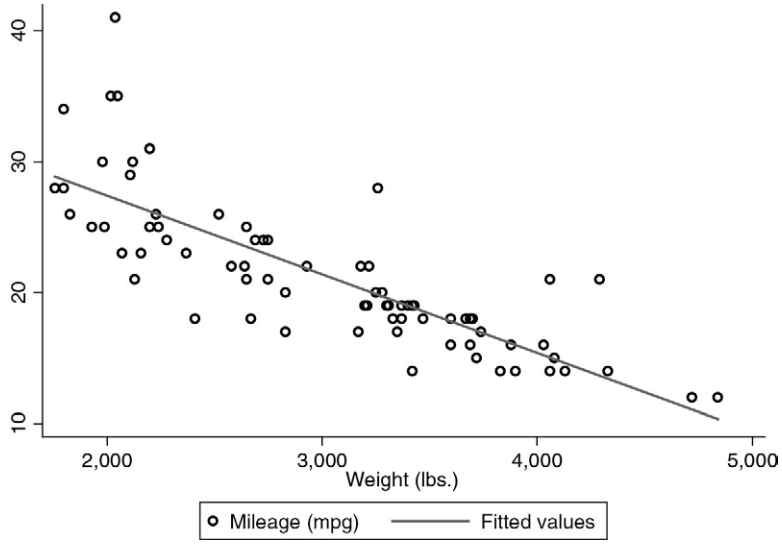


Figura 2.32 ***Gráfico de dispersão entre as variáveis*** **mpg** ***e*** **weight*****, com uma linha de tendência.***

twoway scatter mpg weight || lfit mpg weight

RESULTADOS 2.33 Gerando gráfico de dispersão para duas variáveis, com a linha de tendência.

```
. twoway scatter mpg weight || lfit mpg weight
```

Para acessar os comandos anteriormente apresentados, via barra de menus, devemos selecionar as seguintes opções: *Graphics* → *Twoway graph (scatter, line, etc.)*. Irá surgir a janela da Figura 2.33.

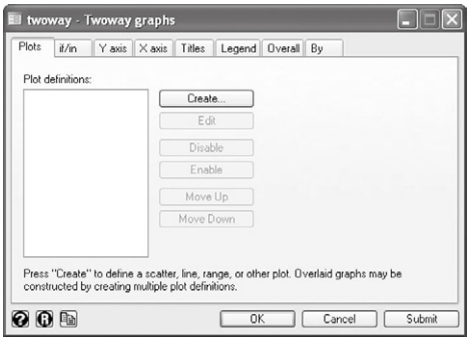


Figura 2.33 *Janela de configuração – Comando* **twoway.**

Basta que cliquemos no botão *Create*, para gerar um novo gráfico. Ao cliclarmos, surgirá outra janela, na qual informaremos o tipo de gráfico e as variáveis a serem utilizadas (Figura 2.34).

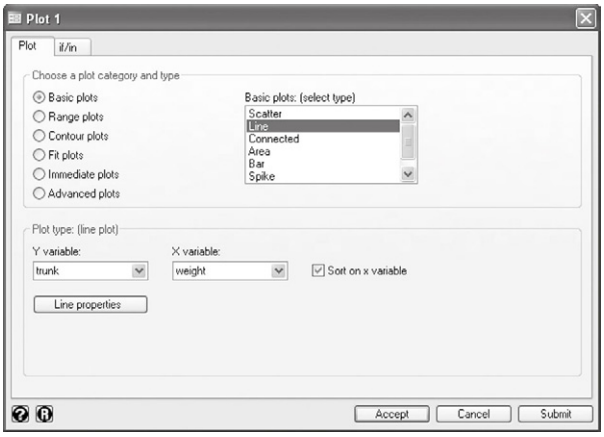


Figura 2.34 *Janela de configuração – Comando* **twoway** *– Criando novo gráfico.*

Ao clicarmos no botão *Accept*, será armazenado o novo gráfico a ser gerado. Assim, poderemos repetir o processo e solicitar quantos gráficos desejamos que o Stata® gere.

O Stata® apresenta diferentes versões do gráfico de barras. O comando **twoway bar** é apenas uma variação do comando que já foi visto.

Também existem gráficos que não fazem parte da família **twoway**. Por exemplo, para gerar um gráfico de barras podemos utilizar o comando **graph bar** (Sintaxe 2.23).

SINTAXE 2.23 Comando graph bar.

graph plot yvars [, over(varname1)]

Em que:

- plot: Representa o gráfico; nessa opção podem ser utilizados: **bar** (barras verticais) e **hbar** (barras horizontais).
- yvars: Lista de variáveis, separadas por espaços em branco.
- over: Opção que indica qual a variável (varname1) que será utilizada para segregar as demais.

Por exemplo, imagine que queremos obter gráficos de barras das variáveis *weight* e *price*, separando-as de acordo com a origem dos veículos (variável *foreign*) (Figura 2.35). Para isso, basta digitarmos o seguinte comando:

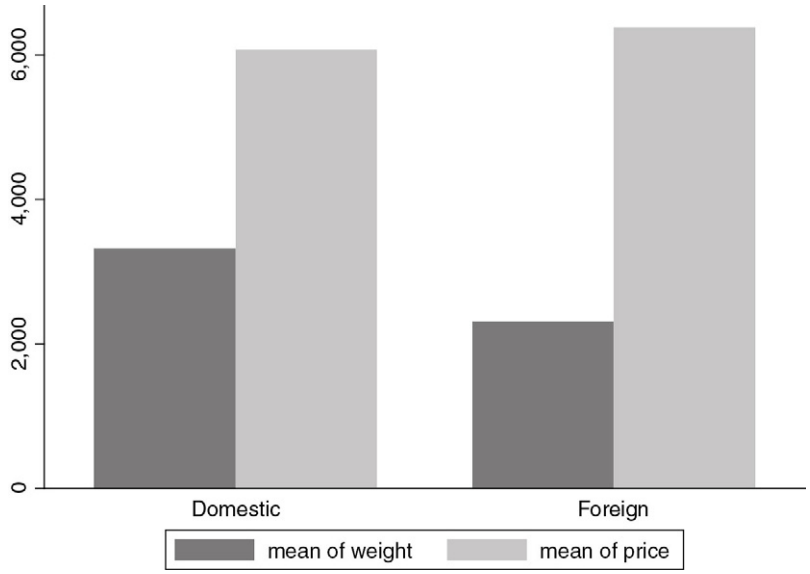


Figura 2.35 ***Gráfico de barras entre as variáveis*** **weight** ***e*** **price,** ***separando os resultados pelas categorias da variável*** **foreign.**

graph bar weight price, over(foreign)

RESULTADOS 2.34 Gerando gráfico de barras para duas variáveis, separando os resultados por outra variável.

```
. graph bar weight price, over(foreign)
```

Para acessar esse comando, por meio da barra de menus, podemos utilizar as seguintes opções: *Graphics* → *Bar chart*. Aparecerá a janela da Figura 2.36.

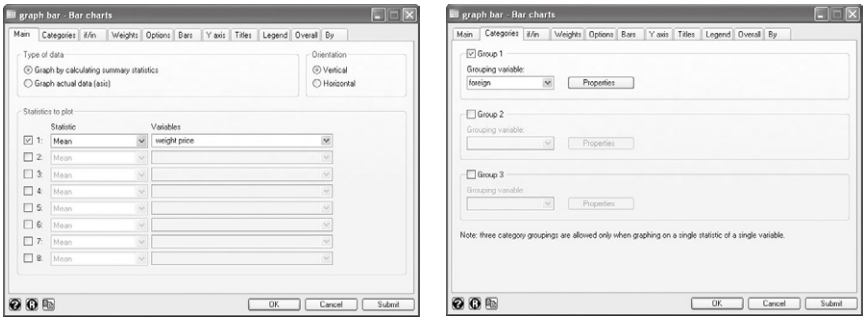


Figura 2.36 *Janela de configuração – Comando* **graph bar.**

2.5. CASO APLICADO

A base de dados **Rentabilidade_MM.dta**, divulgada na revista *Exame Melhores e Maiores*, contém a Rentabilidade Ajustada com data referência de 2007 para mil empresas. Em uma inspeção inicial dos dados, é possível observar a existência de um número significativo de dados faltantes na amostra (Figura 2.37).

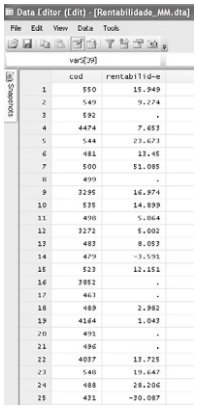


Figura 2.37 *Dados faltantes na base de dados* **Rentabilidade_MM.dta.**

A existência de dados faltantes (*missings*) pode interferir no cálculo de certas estatísticas descritivas desejadas, podendo acarretar um viés na análise dos resultados. Dessa maneira, os dados ausentes foram excluídos da amostra, conforme o seguinte comando:

drop if rentabilidade==.

RESULTADOS 2.35 Apagando valores faltantes (*missings*).

```
. drop if rentabilidade==.
```

Esse procedimento indicou a exclusão inicial de 173 empresas, resultando em uma amostra inicial de análise de 827 empresas. Diante das considerações iniciais expostas, o comando **summarize** do Stata® foi utilizado para que um primeiro diagnóstico sobre a amostra pudesse ser realizado.

summarize rentabilidade, detail

RESULTADOS 2.36 Estatísticas descritivas detalhadas da variável *rentabilidade*.

```
. summarize rentabilidade, detail

                        Rentabilidade Ajustada
-------------------------------------------------------------
      Percentiles      Smallest
 1%     -161.448       -988.895
 5%      -16.607       -680.048
10%       -4.286       -541.027       Obs                 827
25%         3.35        -301.31       Sum of Wgt.         827

50%       10.377                      Mean           6.629724
                       Largest        Std. Dev.      54.38093
75%       19.475        73.138
90%       30.334        74.538       Variance       2957.286
95%       41.249        94.279       Skewness      -11.77969
99%       61.512       100.023       Kurtosis       182.6396
```

Em que:
Mean = Média
Std. Dev. = Desvio-padrão

Variance = Variância
Skewness = Assimetria
Kurtosis = Curtose
Percentiles = Percentís
Mediana = Percentis 50%

Por intermédio das medidas de posição é possível avaliar onde os dados estão concentrados, possibilitando detectar quais são, aparentemente, os valores típicos ou centrais. Calculando as estatísticas descritivas, obteve-se uma média de 6,63 e mediana de 10,38. Uma vez que a média é inferior à mediana calculada, uma primeira conclusão a ser alcançada seria a de que valores extremamente baixos interferiram no cálculo da média, "puxando-a para baixo". Essa hipótese é corroborada pelos valores máximos e mínimos encontrados (percentil 99%: 100,02; percentil 1%: -988,90). O percentil 1% de -988,90 demonstra um comportamento bem destoante do comportamento médio da amostra.

Entretanto, a análise das medidas de tendência central por si só não permite um entendimento completo, impossibilitando avaliar a regularidade com a qual as observações se apresentam. Para se estimar a variação existente nos dados, isto é, como os mesmos estão espalhados, mostra-se necessário o cálculo de medidas tais como a variância e o desvio-padrão. A variância e o desvio-padrão calculados para a amostra foram de 2957,29 e de 54,38, respectivamente. O desvio-padrão nada mais é do que a raiz quadrada da variância, transformando a medida de acordo com a unidade original dos dados. O coeficiente de variação, por sua vez, fornece meios adicionais para a interpretação da magnitude do desvio-padrão: seu cálculo demonstrou um patamar de variação das observações de cerca de 820% (54,38/6,63*100); valor este extremamente elevado, o que caracteriza uma alta dispersão dos dados.

O fato de ter sido encontrada uma média inferior à mediana denota uma assimetria na distribuição dos dados, mais especificamente à esquerda (negativa), constatação corroborada pelo coeficiente de assimetria de -11,80. Por fim, o quarto momento da distribuição, isto é, a curtose, indicou se tratar de uma distribuição leptocúrtica, uma vez que o coeficiente de curtose foi superior a 0 (180,74). O pico mais pronunciado e a cauda longa apontada para a direita podem ser observados no histograma esboçado ao se digitar o seguinte comando: (Figura 2.38)

histogram rentabilidade

RESULTADOS 2.37 Histograma.

```
. histogram rentabilidade
(bin=28, start=-988.89502, width=38.889929)
```

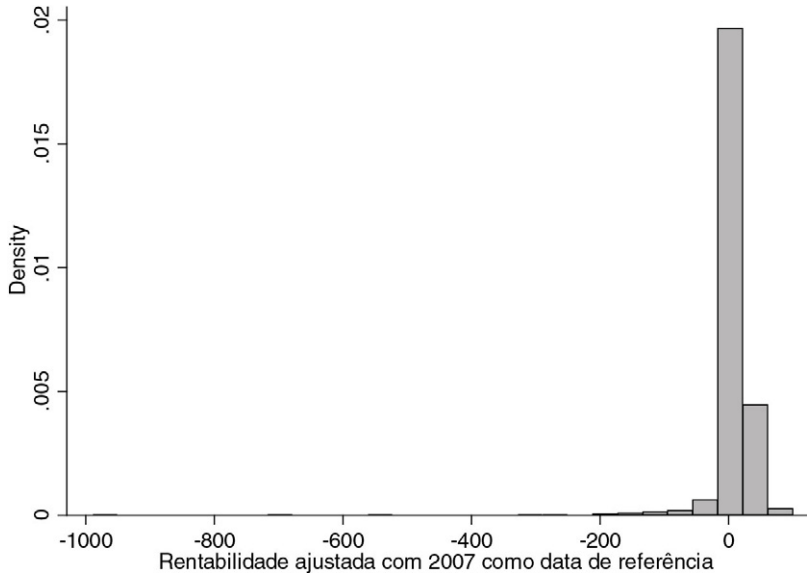


Figura 2.38 ***Histograma da variável*** **rentabilidade.**

Esse critério resultou na exclusão de oito empresas. Outra maneira apresentada por Stevenson (1981) utiliza-se do diagrama box-plot e do cálculo do intervalo interquartil:

$$\left[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1),\right] \qquad \text{[Equação 2.2]}$$

Os quartis são medidas de posição que segregam um conjunto de dados, dispostos em ordem crescente em quatro partes com dimensões iguais, em que o 1° quartil (Q_1 ou 25° percentil) significa que 25% dos dados são inferiores a Q_1 ou que 75% dos dados são superiores a Q_1, o 2° quartil (Q_2 ou 50° percentil) corresponde a mediana e significa, como discutido, que 50% dos dados são inferiores a Q_2, e o terceiro quartil (Q_3 ou 75° percentil) significa que 75% dos dados são inferiores a Q_3 ou que 25% dos dados são superiores a Q_3.

Segundo informações apresentadas na estatística descritiva, o primeiro e o terceiro quartis equivalem a 3,35 e 19,475, respectivamente, resultando em um intervalo interquartil (Q3-Q1) de 16,125. Aplicando a Equação 2.2, com k = 1,5, constata-se que devem ser excluídos valores abaixo de -20,84 e valores acima de 43,66. A exclusão pode ser realizada a partir do seguinte comando:

drop if rentabilidade <=-20.84 | rentabilidade>=43.66

RESULTADOS 2.38 Excluindo observações consideradas *outliers*.

```
. drop if rentabilidade <=-20.84 | rentabilidade >= 43.66
```

Por esse método, 70 empresas foram excluídas da amostra. Levando em conta os dados finais após exclusão dos *outliers*, as estatísticas descritivas foram elaboradas novamente, conforme apresentado nos Resultados 2.39.

summarize rentabilidade, detail

RESULTADOS 2.39 Estatísticas descritivas detalhadas.

```
. summarize rentabilidade, detail

                          Rentabilidade Ajustada
-------------------------------------------------------------
      Percentiles      Smallest
 1%     -16.598         -20.644
 5%      -6.298         -19.811
10%      -1.225         -19.728       Obs                 757
25%       3.977         -19.283       Sum of Wgt.         757

50%      10.377                       Mean           11.40703
                        Largest       Std. Dev.      11.30906
75%      18.334          42.832
90%      26.985          43.059       Variance       127.8947
95%      32.541           43.22       Skewness       .2213965
99%      41.003          43.322       Kurtosis       3.281288
```

A mediana, que antes se encontrava no patamar de 10,38, após a exclusão dos *outliers* permaneceu a mesma. Contudo, a média, antes influenciada por valores extremos, aproximou-se da mediana, passando de 6,63 para 11,41. A assimetria, que antes era à esquerda (negativa), com um coeficiente de assimetria de -11,80 (e uma mediana superior à média), passou a ser à direita e bem menos pronunciada (coeficiente de 0,222). Da mesma maneira, houve uma redução significativa no coeficiente da curtose, que passou de 180,74 para 0,291. Portanto, o terceiro e o quarto momentos demonstraram uma aproximação da distribuição à normal, como demonstrado pelo histograma esboçado ao se digitar o seguinte comando (Figura 2.39):

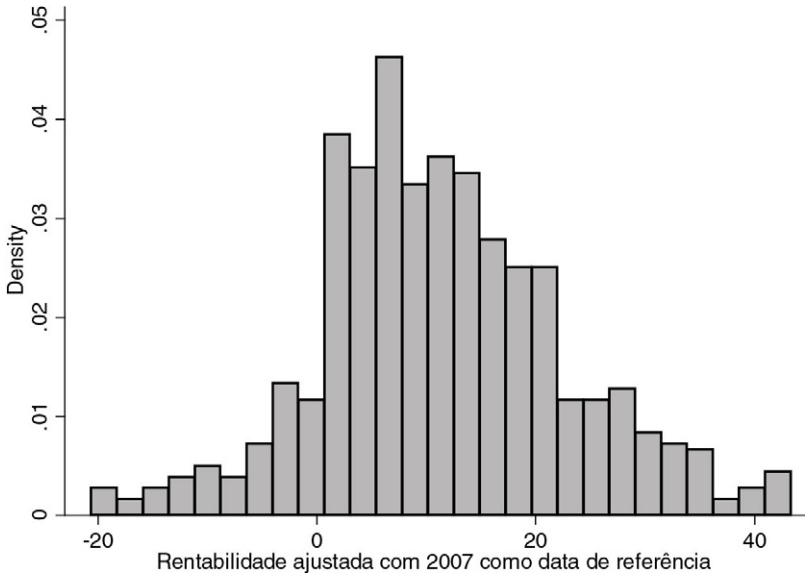


Figura 2.39 ***Novo histograma da variável*** **rentabilidade.**

```
histogram rentabilidade
```

RESULTADOS 2.40 Novo histograma.

```
. histogram rentabilidade
(bin=27, start=-20.643999, width=2.369111)
```

A variância e o desvio-padrão apresentaram uma queda brusca, quando comparados aos valores obtidos na amostra completa (sem a exclusão de *outliers*). Seus valores calculados foram de 127,90 e 11,31, respectivamente. O coeficiente de variação, apesar de ainda relativamente elevado (11,31/11,41 = 99,12%), foi bastante inferior ao coeficiente encontrado anteriormente, de cerca de 820%.

Os resultados explicitam os efeitos que a presença de *outliers* pode ocasionar na estimação das estatísticas descritivas e nas inferências sobre a população subjacente à amostra. Os *outliers* distorceram o cálculo dos quatro momentos da amostra (média, variância, assimetria e curtose), interferindo na distribuição dos dados, afastando-a significativamente da distribuição normal. A não exclusão dessas observações poderia resultar em conclusões errôneas por parte do pesquisador, reduzindo, assim, a possibilidade de generalização de resultados.

2.6. EXERCÍCIOS

1. Inicialmente, solicite a abertura da base de dados **auto.dta** utilizando o comando sysuse (**sysuse auto**). Após a abertura dessa base de dados, calcule as estatísticas descritivas da variável *rep78* (número de reparos no ano de 1978). Pergunta-se:
   a. Qual é o número total de observações?
   b. Qual é o número de *missings* (dados faltantes)?
   c. Qual é o valor mínimo da variável x?
   d. Qual é o valor máximo da variável x?

2. Com a mesma base de dados **auto.dta** utilizada na questão 1, com relação à variável *weight*, pede-se:
   a. Existe algum caso com informações faltantes (*missing*)?
   b. Calcule as seguintes medidas de tendência central: média, mediana e quartis.
   c. Calcule as medidas de dispersão: amplitude, variância, desvio-padrão.
   d. Estime os coeficientes para as seguintes medidas de forma: Assimetria e Curtose.

3. Com a mesma base de dados **auto.dta** utilizado na questão 1, pede-se:
   a. Elabore um histograma desta vez para a variável *gear_ratio* (razão da engrenagem do câmbio). Pode-se afirmar que essa variável se comporta como uma normal? Realize os testes destinados para tal.
   b. Elabore um histograma para a variável *rep78*. Cabe ressaltar que se tratam de dados discretos, devendo essa característica ser especificada quando da elaboração do gráfico.
   c. Elabore um gráfico de dispersão para avaliar se existe uma relação entre o preço (*price*) e a potência dos alto-falantes (*headroom*).

4. A seguir está apresentada a série histórica do IPCA de jan./2010 até dez./2012. Com base nesses dados pede-se:

Índice do mês (em %)					
jan./10	0,75	jan./11	0,83	jan./12	0,56
fev./10	0,78	fev./11	0,80	fev./12	0,45
mar./10	0,52	mar./11	0,79	mar./12	0,21
abr./10	0,57	abr./11	0,77	abr./12	0,64
maio/10	0,43	maio/11	0,47	maio/12	0,36
jun./10	0,00	jun./11	0,15	jun./12	0,08
jul./10	0,01	jul./11	0,16	jul./12	0,43
ago./10	0,04	ago./11	0,37	ago./12	0,41
set./10	0,45	set./11	0,53	set./12	0,57
out./10	0,75	out./11	0,43	out./12	0,59
nov./10	0,83	nov./11	0,52	nov./12	0,60
dez./10	0,63	dez./11	0,50	dez./12	0,79

   a. Elabore um gráfico de dispersão para a série histórica apresentada.
   b. Elabore um gráfico de linha para os mesmos dados.

CAPÍTULO 3

Testes de Hipótese e Análise de Variância (ANOVA)

3.1. INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Frequentemente precisamos obter conclusões válidas sobre um grande grupo de indivíduos ou objetos. Para compreensão de inferência estatística, os dois conceitos mais importantes são: população e amostra. Uma população pode ser definida como a totalidade de todas as observações possíveis sobre medidas ou ocorrências. A população pode ser finita ou infinita.

Contudo, em vez de examinar todo o grupo (população), pode-se estudar apenas uma pequena parte (amostra) dessa população. Desde que essa amostra seja representativa dessa população, podemos fazer inferências sobre a segunda, a partir da análise da primeira. A inferência estatística é o processo que tem por objetivo inferir (generalizar) determinados fatos acerca da população, a partir de resultados observados na amostra.

3.2. TESTES DE HIPÓTESE COM UMA AMOSTRA

Nesse tipo de teste, geralmente é feita uma afirmação sobre a média populacional, e depois a comparamos com a estatística obtida a partir da amostra (FÁVERO *et al.*, 2009). Para isso, quando os dados possuem distribuição normal utilizamos a estatística t, que segue uma distribuição t de Student com n-1 g.l. (graus de liberdade). A estatística t é calculada a partir da média amostral, da média populacional, do desvio-padrão amostral e do tamanho da amostra, conforme demonstrado na expressão a seguir:

$$t = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \qquad \text{[Equação 3.1]}$$

Suponha que em uma pesquisa anual com o histórico de 10 anos a média de reparação dos carros tenha se apresentado constante nos últimos anos. Não existem indícios de que essa média foi alterada no ano de análise. Contudo, por meio do teste de hipóteses buscam-se alterações nesse valor médio. O teste de hipóteses que você deseja implementar visa verificar se a média do valor de vendas está aumentando ou diminuindo.

1. O primeiro passo do teste consiste em enunciar a hipótese nula (H_0) e a hipótese alternativa (H_1) do teste, que no caso são:

$$H_0 : \mu = 3$$
$$H_1 : \mu \neq 3$$

2. O segundo passo consiste na definição do nível de significância estatística do teste (α), geralmente igual a 5% em ciências sociais aplicadas.

3. Dado que o tamanho da amostra é n = 69, teremos n-1 g.l. = 68 g.l. para o teste t.

No Stata® o teste t para a comparação de uma média é realizado por meio do comando **ttest**. No caso anteriormente apresentado será utilizada a base de dados **auto.dta**, que acompanha o aplicativo (lembre-se de que a mesma poderá ser aberta com o comando **sysuse** auto), sendo a nossa variável de interesse a *rep78* (reparação). Para isso, podemos utilizar o seguinte comando:

ttest rep78 = =3

A Figura 3.1 apresenta o passo a passo para a elaboração do teste por meio das janelas de comando.

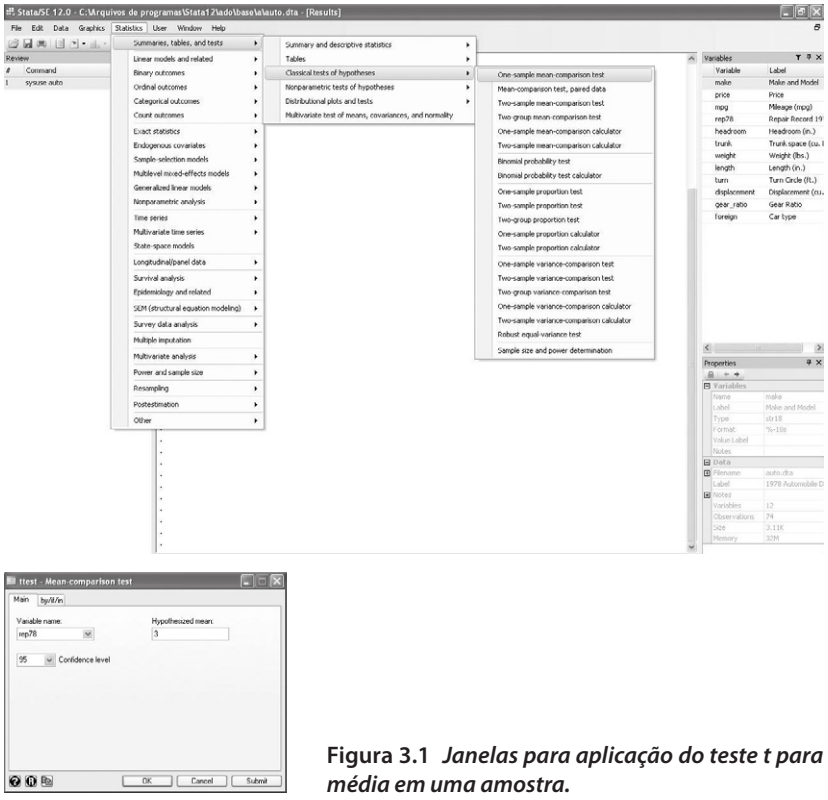


Figura 3.1 ***Janelas para aplicação do teste t para média em uma amostra.***

RESULTADOS 3.1 Teste t para uma amostra.

```
. ttest rep78==3

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   rep78 |      69    3.405797    .1191738    .9899323    3.167989    3.643605
------------------------------------------------------------------------------
    mean = mean(rep78)                                            t =   3.4051
Ho: mean = 3                                     degrees of freedom =       68

    Ha: mean < 3                 Ha: mean != 3                 Ha: mean > 3
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0011          Pr(T > t) = 0.0006
```

O método de construção de um teste de hipóteses parte da fixação do nível de significância α. Os resultados do teste t são analisados a partir da comparação entre o nível de significância e a probabilidade ou p-valor do teste ou da comparação entre a estatística t calculada e o respectivo valor crítico para o nível de significância definido.

O p-valor pode ser usado para tomar decisões em um teste de hipóteses, observando-se que:

1. Se o p-valor é menor que α, o valor da estatística de teste está na região de rejeição da hipótese nula.
2. Se o p-valor é maior ou igual a α, o valor da estatística de teste não está na região de rejeição da hipótese nula, ou seja, na região crítica do teste (RC).

   Portanto, deve-se rejeitar H_0 se o p-valor $< \alpha$.

No exemplo anterior, podemos notar que os resultados exibidos pelo Stata® apresentam p-valores para três hipóteses alternativas, enquanto a hipótese nula é a mesma H_0: $\mu = 3$. Nossa hipótese alternativa foi de que H_1: $\mu \neq 3$. Considerando essas hipóteses, o teste retornou um p-valor de 0,0011 (ou 0,11%), que é inferior ao nível de significância fixado (0,05 ou 5%) e conduz à rejeição da hipótese nula de que a média de reparos anual seria igual a três.

Caso desejássemos saber se a média seria igual ou inferior a três, como hipóteses nula e alternativa teríamos H_0: $\mu = 3$ e H_1: $\mu < 3$, respectivamente, e verificaríamos que, com uma probabilidade de 0,9994, a média seria estatisticamente igual a três. Todavia, caso as hipóteses nula e alternativa fossem H_0: $\mu = 3$ e H_1: $\mu > 3$, respectivamente, veríamos que o teste resultou em um p-valor de 0,0006, o que levaria à aceitação da hipótese alternativa de que a média seria maior do que três.

De acordo com Levine *et al.* (2000):

- O teste t de uma amostra é considerado um procedimento paramétrico clássico.
- Como tal, estabelece uma série de pressupostos restritivos que devem se manter, se quisermos estar seguros de que os resultados que obtivermos ao empregar o teste são válidos.

- Em particular, para utilizar o teste t para uma amostra, pressupõe-se que os dados numéricos obtidos são extraídos independentemente e representam uma amostra aleatória de uma população que é normalmente distribuída, ou seja, deve-se seguir uma distribuição normal.

3.3. TESTES DE HIPÓTESE COM DUAS AMOSTRAS

Em diversas situações estaremos interessados em verificar se as médias de duas amostras apresentam diferenças significativas ou se podem ser consideradas como estatisticamente iguais.

Para esse fim, deve-se lançar mão de testes apropriados para essas comparações.

O caso mais abrangente é aquele em que existem populações com variâncias desiguais. Para esse caso, é necessário calcular os graus de liberdade da distribuição t, considerando as variâncias de ambas as amostras. Em sentido contrário, caso as variâncias fossem iguais, a distribuição t utilizada possuiria n-2 g.l.

Para se testar se as médias das duas populações são estatisticamente diferentes, deve-se usar a seguinte estatística t:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n}\frac{\sigma_2^2}{m}}} \sim T_{k(g.l.)}$$

$$k = \frac{\left(\frac{\sigma_1^2}{n}\frac{\sigma_2^2}{m}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n}\right)^2}{(n-1)} + \frac{\left(\frac{\sigma_2^2}{m}\right)^2}{(m-1)}} \qquad \text{[Equação 3.2]}$$

Voltaremos a utilizar o comando **ttest**. Suponha que desejamos saber se há diferenças entre a média de reparo (*rep78*) dos carros nacionais e estrangeiros (sendo o tipo identificado na variável *foreign*). Considerando o caso mais comum, de que os grupos apresentam variâncias desiguais, utilizaremos o seguinte comando:

ttest rep78, by(foreign) unequal

A Figura 3.2 apresenta o passo a passo para a elaboração do teste por meio das janelas de comando.

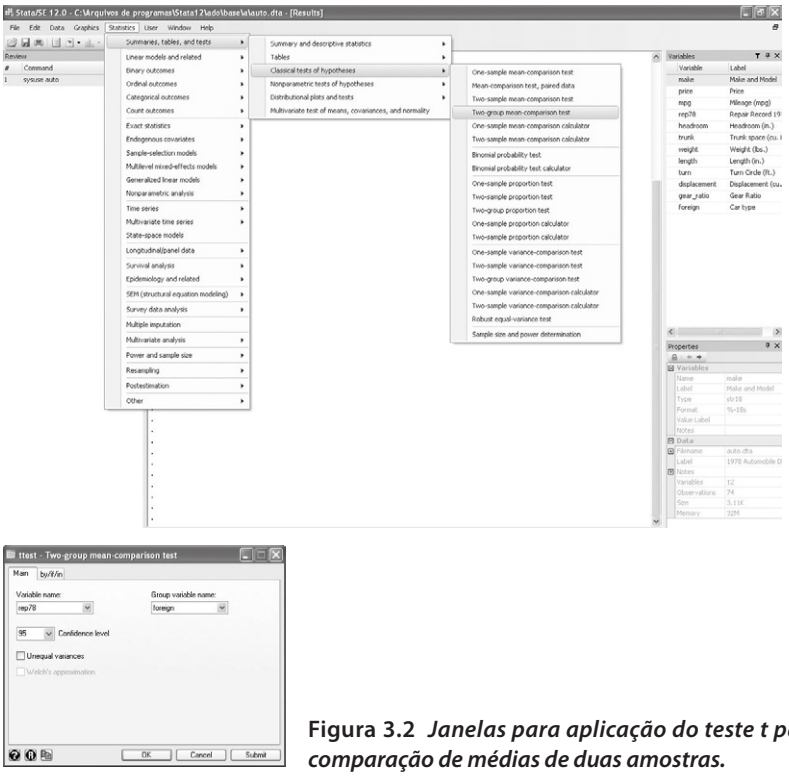


Figura 3.2 ***Janelas para aplicação do teste t para comparação de médias de duas amostras.***

RESULTADOS 3.2 Teste t para duas amostras com variâncias desiguais.

```
. ttest rep78, by(foreign) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      48    3.020833    .1209067     .837666      2.7776    3.264066
 Foreign |      21    4.285714    .1564922    .7171372    3.959277    4.612151
---------+--------------------------------------------------------------------
combined |      69    3.405797    .1191738    .9899323    3.167989    3.643605
---------+--------------------------------------------------------------------
    diff |           -1.264881     .197758               -1.663363   -.8663991
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -6.3961
Ho: diff = 0                     Satterthwaite's degrees of freedom =   44.288

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

De acordo com os p-valores apresentados para cada par de hipóteses nula e alternativa, verifica-se que as médias dos carros nacionais e dos estrangeiros são estatisticamente diferentes (H_0: diff = 0 *versus* H_1: diff ≠ 0, em que diff = média nacionais – média estrangeiros) e que a média dos carros nacionais é menor do que a média de reparos dos estrangeiros (H_0: diff = 0 *versus* H_1: diff < 0).

Caso as variâncias dos grupos fossem iguais, o comando utilizado seria o seguinte:

ttest rep78, by(foreign)

RESULTADOS 3.3 Teste t para duas amostras com variâncias iguais.

```
. ttest rep78, by(foreign)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      48    3.020833    .1209067    .837666     2.7776    3.264066
 Foreign |      21    4.285714    .1564922   .7171372   3.959277    4.612151
---------+--------------------------------------------------------------------
combined |      69    3.405797    .1191738   .9899323   3.167989    3.643605
---------+--------------------------------------------------------------------
    diff |           -1.264881    .2102445              -1.684531   -.8452312
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -6.0162
Ho: diff = 0                                     degrees of freedom =       67

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Nos Resultados 3.3 verifica-se que não houve alterações em relação ao caso anterior.

Qualquer que seja a decisão tomada, estamos sujeitos a cometer erros. Desta maneira, temos:

Erro do Tipo I: rejeitar a hipótese nula quando essa é verdadeira.

α = P(erro do tipo I) = P(rejeitar H_0 | H_0 é verdadeira)

Erro do Tipo II: não rejeitar H_0 quando H_0 é falsa.

β = P(erro do tipo II) = P(não rejeitar H_0 | H_0 é falsa)

Para ser capaz de utilizar o teste t, é necessário determinar se as duas populações (ou amostras) têm a mesma variância, ou não. Nesse caso, utiliza-se o teste F, que compara a variância de duas populações. Nesse caso, as hipóteses nula e alternativa são, respectivamente:

H_0: As duas populações têm a mesma variância.

H_1: As duas populações não têm a mesma variância.

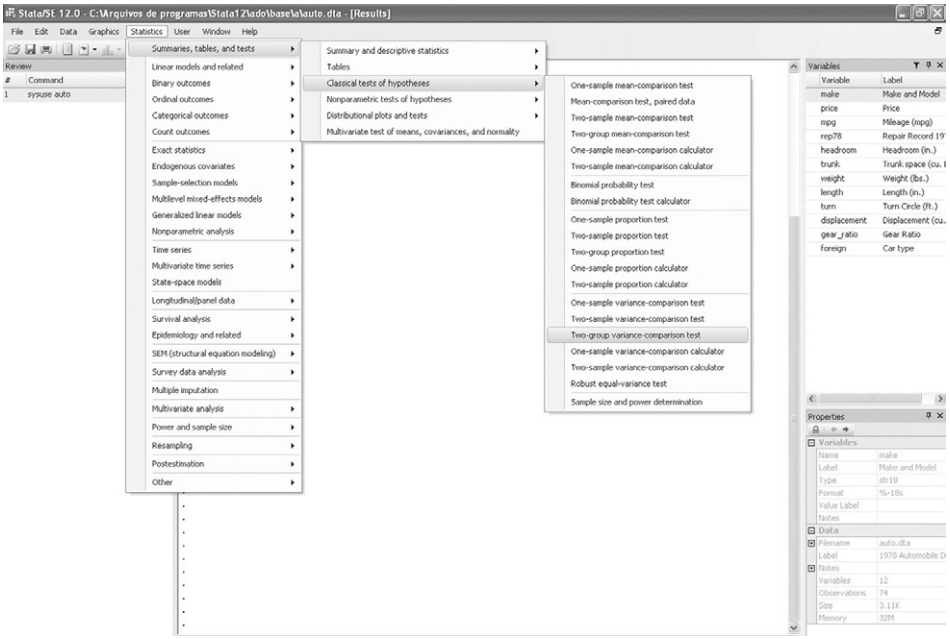
Ao se analisar o resultado do teste F, pode-se determinar se deve ser selecionada a opção *Unequal variances,* ou não. Essa decisão é baseada no teste F, que irá avaliar a variância de duas populações.

Considerando o exemplo anterior, o comando para a execução do teste F para verificar a igualdade (homogeneidade) das variâncias é o seguinte:

sdtest rep78, by(foreign)

A Figura 3.3 apresenta o passo a passo para a elaboração do teste por meio das janelas de comando.

A parte superior dos Resultados 3.4 contém algumas estatísticas descritivas dos dois grupos. Na segunda parte do *output*, é apresentado o teste F propriamente dito. Um p -valor maior igual a 0,05 significa que a hipótese nula, que assume que as variâncias são equivalentes, é aceitável e, portanto, pode-se utilizar a opção padrão (*default*) do programa de variâncias equivalentes do teste t, anteriormente apresentado. Um p-valor menor



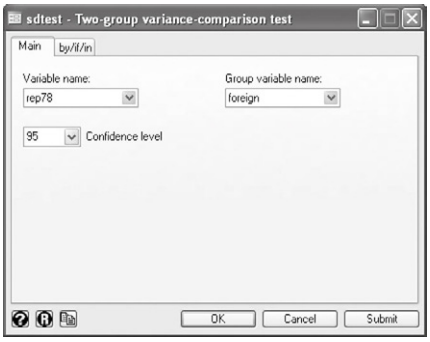


Figura 3.3 ***Janelas para aplicação do teste F para comparação de variâncias de duas amostras.***

RESULTADOS 3.4 Teste F para igualdade de variâncias.

```
. sdtest rep78, by(foreign)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      48    3.020833    .1209067    .837666      2.7776    3.264066
 Foreign |      21    4.285714    .1564922    .7171372    3.959277    4.612151
---------+--------------------------------------------------------------------
combined |      69    3.405797    .1191738    .9899323    3.167989    3.643605
------------------------------------------------------------------------------
    ratio = sd(Domestic) / sd(Foreign)                             f =   1.3644
Ho: ratio = 1                                  degrees of freedom =   47, 20

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.7726         2*Pr(F > f) = 0.4548          Pr(F > f) = 0.2274
```

que 0,05 significa que é necessário selecionar a opção *Unequal variances* ao realizar o teste t. Nesse caso, o nível de significância é confortavelmente acima de 0,05, e portanto variâncias equivalentes são assumidas (Figura 3.4).

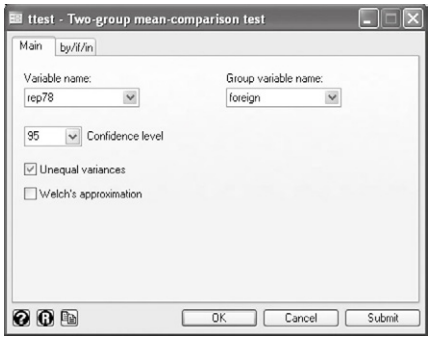


Figura 3.4 ***Seleção da opção*** **Unequal variances.**

3.4. ANÁLISE DE VARIÂNCIA (ANOVA)

A análise de variância (ANOVA) engloba um grupo de métodos para testar hipóteses sobre diferenças entre médias. O grupo de aplicações alcança desde uma simples análise em que se compara a média da variável y ao longo das categorias da variável x, até situações mais complexas, com múltiplas categorias e medidas para a variável x. O teste t para hipóteses relacionadas a uma única média (*one sample*) ou a um par de médias (*two samples*) corresponde às formas elementares da ANOVA.

Testes baseados em postos (*rank tests*) não paramétricos, incluindo o teste de sinais, Mann-Whitney e Kruskal-Wallis, empregam uma diferente abordagem para comparar distribuições. Esses testes assumem pressupostos mais fracos sobre a medida, o formato e a dispersão da distribuição. Consequentemente, eles permanecem válidos sob um grupo mais amplo de condições do que a ANOVA e seus testes similares "paramétricos". Analistas cuidadosos muitas vezes empregam os testes paramétricos e não paramétricos em conjunto, checando para avaliar se ambos apontam a mesma conclusão.

O modelo da ANOVA possui uma flexibilidade considerável, englobando um amplo grupo de modelos. A ANOVA pode se ajustar para *one-way*, n-*way* e a análise de covariância (ANCOVA) para dados balanceados e não balanceados (quando há dados faltantes). Uma característica importante do Stata® é que ele não tem modos ou módulos. Não é necessário instalar um módulo específico para estimar um modelo ANOVA, basta digitar o comando. Essa característica possibilita que outros comandos Stata® sejam intercalados, levando a um melhor entendimento dos dados.

3.5. ANÁLISE MULTIVARIADA DE VARIÂNCIA

Suponha que uma instituição financeira estivesse interessada em investigar a adequação do limite de crédito concedido aos clientes de uma carteira específica. Para tanto, faz uso da análise da relação entre o valor tomado e o limite de crédito nos produtos

de cheque especial e de cartão de crédito. Além disso, imagine que a empresa esteja interessada em analisar se existem diferenças significativas para esse quesito em relação às classes sociais dos clientes.

Para tanto, as variáveis dependentes são os percentuais de utilização do crédito em relação aos respectivos limites concedidos no cheque especial e no cartão de crédito e a variável independente refere-se às classes sociais.

Assim, a hipótese nula pode ser descrita da seguinte maneira (FÁVERO *et al.*, 2009):

$$H_0 : \begin{pmatrix} \mu_{cartão,classeA} \\ \mu_{cheque,classeA} \end{pmatrix} = \begin{pmatrix} \mu_{cartão,classeB} \\ \mu_{cheque,classeB} \end{pmatrix} = \begin{pmatrix} \mu_{cartão,classeC} \\ \mu_{cheque,classeC} \end{pmatrix} \quad \text{[Equação 3.3]}$$

Os dados utilizados no exemplo estão disponibilizados no arquivo **exemplomanova.dta**.

Antes de realizarmos a MANOVA propriamente dita, é necessário que averiguemos a validade dos pressupostos subjacentes à utilização dessa técnica.

Uma das suposições estabelecidas pela MANOVA é de que as variáveis sejam provenientes de um grupo de populações que seguem uma distribuição normal multivariada. Isso significa que cada uma das variáveis dependentes é normalmente distribuída dentro do grupo, que qualquer combinação linear das variáveis dependentes é normalmente distribuída, e que todos os subconjuntos das variáveis devem seguir uma distribuição normal multivariada. Um teste para verificação desta hipótese pode ser aplicado usando-se o comando **mvtest normality**, que foi introduzido no Stata® versão 11. No nosso exemplo, o teste pode ser realizado por meio do seguinte comando:

mvtest normality perc_cartao perc_cheque

RESULTADOS 3.5 Teste de normalidade multivariada.

```
. mvtest normality perc_cartao perc_cheque

Test for multivariate normality

    Doornik-Hansen                chi2(4) =   12.736   Prob>chi2 =  0.0126
```

Com respeito ao Erro do tipo I, apesar do teste de normalidade multivariada de Doornik-Hansen rejeitar a hipótese nula sobre a existência de normalidade multivariada das variáveis selecionadas, a um nível de significância de 5%, a MANOVA tende a ser robusta a pequenas violações da suposição de normalidade multivariada (http://www.ats.ucla.edu/stat/stata/dae/manova1.htm, acesso em 10/04/2013).

Além da premissa de normalidade multivariada das variáveis dependentes, a MANOVA pressupõe igualdade de suas matrizes de variância-covariância, as quais são avaliadas pelo teste Box's M e pelo teste de Levene. O teste Box's M é utilizado para investigar se há indícios que levam à rejeição da hipótese nula de igualdade das matrizes de variância-covariância entre os grupos, tendo em vista que há mais de uma variável dependente no estudo. O comando geral para esse teste pode ser dado por:

mvtest covariance perc_cartao perc_cheque, by(classesocial)

A Figura 3.5 apresenta o passo a passo para a elaboração do teste por meio das janelas de comando.

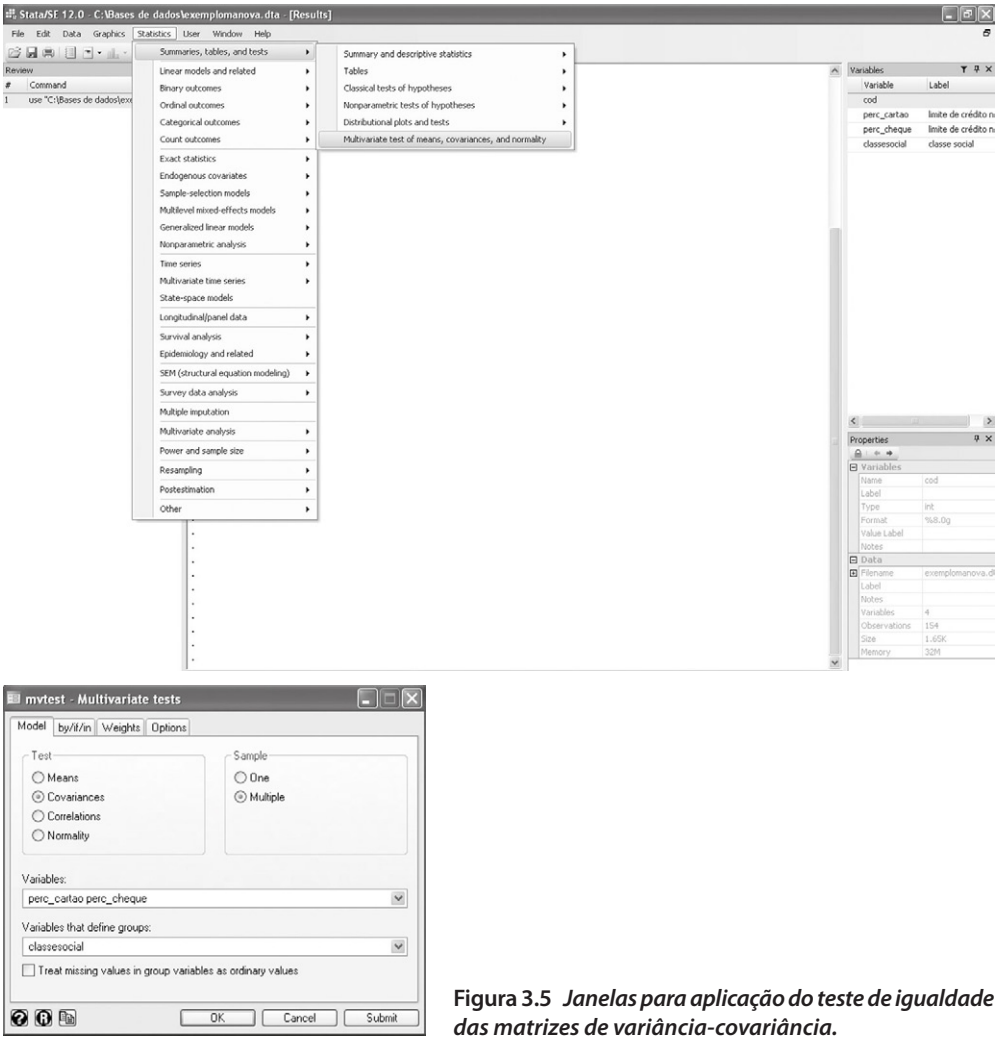


Figura 3.5 ***Janelas para aplicação do teste de igualdade das matrizes de variância-covariância.***

RESULTADOS 3.6 Teste de igualdade das matrizes de variância-covariância.

```
. mvtest covariance perc_cartao perc_cheque, by( classesocial)

Test of equality of covariance matrices across 3 samples

      Modified LR chi2 =  6.443816
      Box F(6,68119.7) =      1.05      Prob > F =  0.3917
           Box chi2(6) =      6.29   Prob > chi2 =  0.3916
```

Os resultados do teste Box's M sugerem a não rejeição da hipótese nula de igualdade das matrizes de covariância, com significância de 5%.

O teste de Levene, por sua vez, é utilizado para analisar a existência de homogeneidade em cada variável dependente individualmente (FÁVERO *et al.*, 2009). O comando **robvar** estima o teste de Levene de igualdade de variâncias (denominado W0). Os comandos para analisar a homogeneidade de variância das variáveis *perc_cartao* e *perc_cheque* são, respectivamente, apresentados a seguir:

robvar perc_cartao, by(classesocial)

robvar perc_cheque, by(classesocial)

As Figuras 3.6 e 3.7 apresentam o passo a passo para elaboração do teste para cada variável por meio das janelas de comando.

RESULTADOS 3.7 Teste de Levene para a variável *perc_cartao*.

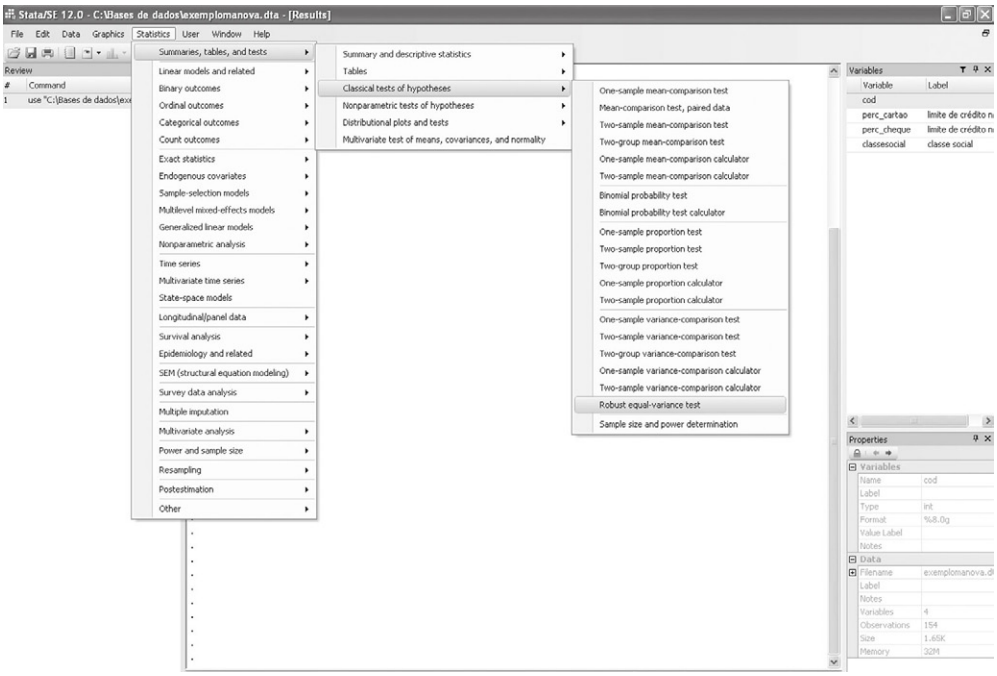
```
. robvar perc_cartao, by(classesocial)

classesocia |        Summary of perc_cartao
          l |         Mean   Std. Dev.       Freq.
------------+------------------------------------
          a |   .46428525   .20161313          27
          b |   .48449666   .29820705          46
          c |   .53708524   .27325176          81
------------+------------------------------------
      Total |   .50861333   .27028637         154

W0   =  4.4310598   df(2, 151)     Pr > F = 0.01348867

W50  =  3.9457515   df(2, 151)     Pr > F = 0.02136283

W10  =  4.3764826   df(2, 151)     Pr > F = 0.01420252
```



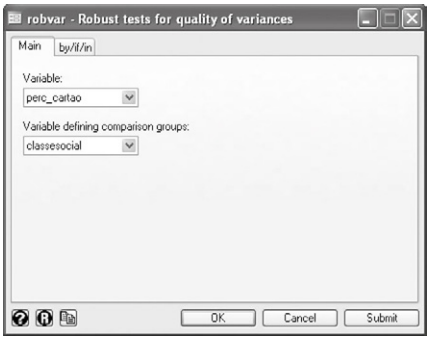


Figura 3.6 ***Janelas para aplicação do teste de Levene para a variável*** **perc_cartao.**

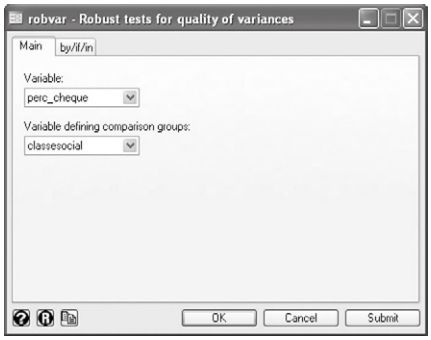


Figura 3.7 ***Janela para aplicação do teste de Levene para a variável*** **perc_cheque.**

RESULTADOS 3.8 Teste de Levene para a variável *perc_cheque*.

```
. robvar perc_cheque, by(classesocial)

classesocia |        Summary of perc_cheque
          l |         Mean   Std. Dev.       Freq.
------------+------------------------------------
          a |   .46132429   .18879981          27
          b |   .52193671   .21979555          46
          c |   .49294528   .20151509          81
------------+------------------------------------
      Total |   .49606112   .20474063         154

W0   =  1.1281371   df(2, 151)      Pr > F = 0.32634777

W50  =  1.2081072   df(2, 151)      Pr > F = 0.30163325

W10  =  1.1449259   df(2, 151)      Pr > F = 0.32099441
```

O resultado do teste de Levene, por sua vez, indica, com nível de significância de 5%, que apenas o percentual de utilização do limite de crédito do cheque especial atende ao pressuposto da homogeneidade de variância. Ou seja, a outra variável dependente (*perc_cartao*) somente observa esse pressuposto se o nível de significância for 1%. Neste sentido, caberá ao pesquisador avaliar o nível de significância a ser adotado no estudo e os respectivos impactos. Para fins didáticos, e tendo em vista os resultados do teste de Box's M, será dada sequência à análise dos outros resultados.

Para a obtenção dos resultados dos testes de médias (Pillai's Trace, Wilks' Lambda, Hotelling's Trace e Roy's Largest Root), por sua vez, basta digitar o seguinte comando:

manova perc_cartao perc_cheque = classesocial

A Figura 3.8 apresenta o passo a passo para elaboração dos testes por meio das janelas de comando.

RESULTADOS 3.9 Testes de médias.

```
. manova perc_cartao perc_cheque = classesocial

                          Number of obs =     154

                          W = Wilks' lambda      L = Lawley-Hotelling trace
                          P = Pillai's trace     R = Roy's largest root

                Source |  Statistic     df   F(df1,     df2) =   F   Prob>F
           ------------+-----------------------------------------------------
           classesoc~l | W   0.9771      2     4.0   300.0      0.87 0.4806 e
                       | P   0.0230            4.0   302.0      0.88 0.4771 a
                       | L   0.0233            4.0   298.0      0.87 0.4841 a
                       | R   0.0132            2.0   151.0      1.00 0.3702 u
                       |-----------------------------------------------------
                  Residual |               151
           ------------+-----------------------------------------------------
                 Total |               153
           ---------------------------------------------------------------------
                          e = exact, a = approximate, u = upper bound on F
```

Figura 3.8 ***Janelas para aplicação dos testes de médias (Pillai's Trace, Wilks' Lambda, Hotelling's Trace e Roy's Largest Root).***

Os testes de médias (Pillai's Trace, Wilks' Lambda, Hotelling's Trace e Roy's Largest Root) sugerem a não rejeição da hipótese nula de igualdade de médias entre as classes sociais em relação aos percentuais de utilização do limite de crédito concedido no cartão de crédito e no cheque especial, indicando adequação da política de crédito da instituição financeira em relação ao não beneficiamento de qualquer classe social em detrimento de outras.

Os resultados apresentados a seguir são coerentes com o que já foi discutido, apontando para a não existência de elementos que levem à rejeição da hipótese nula de igualdade de médias, com nível de significância de 5%, entre as classes sociais.

```
foreach vname in perc_cartao perc_cheque {
anova 'vname' classesocial
}
```

RESULTADOS 3.10 ANOVA - Teste F para a variável *classesocial*.

```
. foreach vname in perc_cartao perc_cheque {
  2.
. anova `vname' classesocial
  3.
. }

                           Number of obs =       154     R-squared     =  0.0130
                           Root MSE      = .270294     Adj R-squared = -0.0001

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+------------------------------------------------------
                   Model |  .145471268     2  .072735634        1.00     0.3719
                         |
              classesoc~l |  .145471268     2  .072735634        1.00     0.3719
                         |
                Residual |  11.0319013   151  .073058949
              -----------+------------------------------------------------------
                   Total |  11.1773725   153  .073054722

                           Number of obs =       154     R-squared     =  0.0100
                           Root MSE      = .205059     Adj R-squared = -0.0031

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+------------------------------------------------------
                   Model |  .064164975     2  .032082487        0.76     0.4681
                         |
              classesoc~l |  .064164975     2  .032082487        0.76     0.4681
                         |
                Residual |  6.34939981   151  .042049005
              -----------+------------------------------------------------------
                   Total |  6.41356479   153  .041918724
```

Para a realização de testes *post-hoc*, avaliando possíveis diferenças entre os grupos, é necessário utilizar o comando **manovatest, showorder**, para determinar a ordem em que os elementos estão dispostos na matriz. Este comando deve ser aplicado após o comando **manova**. É necessário que se conheça a ordem em que os elementos estão dispostos na matriz, a fim de que seja possível prosseguir com a comparação de médias.

```
manovatest, showorder
```

RESULTADOS 3.11 Definindo a ordem em que os elementos estão na matriz.

```
. manovatest, showorder

 Order of columns in the design matrix
      1: (classesocial==1)
      2: (classesocial==2)
      3: (classesocial==3)
      4: _cons
```

Podemos começar comparando a classe social 1 com a média das classes sociais 2 e 3. A hipótese é que as médias dos dois grupos sejam iguais. O resultado anteriormente apresentado indica que o quarto elemento da matriz é a constante, ou seja, será estabelecido como zero no comando **matrix** a seguir. Uma vez criada a matriz (que denominaremos c1), pode-se utilizar o comando **manovatest** para testá-la.

```
matrix c1 = (2,-1,-1,0)
manovatest, test(c1)
```

RESULTADOS 3.12 Testando a classe social 1 em relação às demais.

```
. matrix c1=(2,-1,-1,0)
. manovatest, test(c1)

 Test constraint
 (1)     2*1.classesocial - 2.classesocial - 3.classesocial = 0

                            W = Wilks' lambda      L = Lawley-Hotelling trace
                            P = Pillai's trace     R = Roy's largest root

                Source |  Statistic     df   F(df1,     df2) =   F   Prob>F
           ------------+-----------------------------------------------------
           manovatest | W   0.9887      1     2.0   150.0     0.86 0.4254 e
                      | P   0.0113            2.0   150.0     0.86 0.4254 e
                      | L   0.0115            2.0   150.0     0.86 0.4254 e
                      | R   0.0115            2.0   150.0     0.86 0.4254 e
                      |-----------------------------------------------------
             Residual |               151
           ------------------------------------------------------------------
                            e = exact, a = approximate, u = upper bound on F
```

Os resultados indicam que a classe social 1 não diferiu significativamente das classes sociais 2 e 3. Poder-se-ia desejar comparar duas classes sociais, tais como a 2 e a 3. Novamente, é necessário que se crie uma nova matriz (chamada de c2 no nosso exemplo) para a realização dessa comparação.

```
matrix c2 = (0,1,-1,0)
manovatest, test(c2)
```

RESULTADOS 3.13 Comparando as classes sociais 2 e 3.

```
. matrix c2=(0,1,-1,0)

.
. manovatest, test(c2)

 Test constraint
 (1)     2.classesocial - 3.classesocial = 0

                            W = Wilks' lambda      L = Lawley-Hotelling trace
                            P = Pillai's trace     R = Roy's largest root

                Source |  Statistic     df   F(df1,     df2) =   F   Prob>F
           ------------+-----------------------------------------------------
           manovatest | W   0.9887      1     2.0   150.0     0.86 0.4264 e
                      | P   0.0113            2.0   150.0     0.86 0.4264 e
                      | L   0.0114            2.0   150.0     0.86 0.4264 e
                      | R   0.0114            2.0   150.0     0.86 0.4264 e
                      |-----------------------------------------------------
             Residual |               151
           ------------------------------------------------------------------
                            e = exact, a = approximate, u = upper bound on F
```

Novamente não se identificaram diferenças significativas entre os grupos analisados. A seguir são apresentadas as médias marginais estimadas.

margins classesocial, predict(equation(perc_cartao))

RESULTADOS 3.14 Médias marginais estimadas para a variável *perc_cartao*.

```
. margins classesocial, predict(equation( perc_cartao))

Adjusted predictions                              Number of obs   =        154

Expression   : Linear prediction, predict(equation( perc_cartao))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
classesocial |
           1 |   .4642853   .0520181     8.93   0.000     .3623316    .5662389
           2 |   .4844967   .0398527    12.16   0.000     .4063868    .5626065
           3 |   .5370852   .0300327    17.88   0.000     .4782223    .5959482
------------------------------------------------------------------------------
```

margins classesocial, predict(equation(perc_cheque))

RESULTADOS 3.15 Médias marginais estimadas para a variável *perc_cheque*.

```
. margins classesocial, predict(equation( perc_cheque))

Adjusted predictions                              Number of obs   =        154

Expression   : Linear prediction, predict(equation( perc_cheque))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
classesocial |
           1 |   .4613243   .0394635    11.69   0.000     .3839772    .5386714
           2 |   .5219367   .0302342    17.26   0.000     .4626787    .5811947
           3 |   .4929453   .0227843    21.64   0.000     .4482889    .5376017
------------------------------------------------------------------------------
```

3.6. EXERCÍCIOS

1. Um investidor possui 13 ativos que lhe renderam os seguintes retornos:

8,4% 4,6% 11,9% 15,3% 7,6% 12,2% 9,0% 15,6% 14,5% 6,0% 18,8% 9,1% 18,1%

Investimentos com perfis de risco semelhantes lhe renderiam cerca de 12%. Dessa maneira, calcule a média da rentabilidade e avalie se está estatisticamente abaixo ou acima da rentabilidade média oferecida pelo mercado.

2. O arquivo **endividamento.dta** contém o endividamento de longo prazo das mil maiores empresas de capital aberto do país para o ano de 2007, sendo estas empresas segregadas em três ramos de atuação (comércio, indústria ou serviços). Com base nesse arquivo, responda as seguintes questões:
   - **a.** Existem dados faltantes? Exclua esses casos.
   - **b.** Qual a média do endividamento?
   - **c.** Teste a hipótese de que a média da variável *endividamento_lp* é igual a 20% a partir de um teste bicaudal. Reporte o p-valor. Devemos rejeitar a hipótese a um nível de 5% de significância?
   - **d.** A variável *ramo_atividade* contém informações sobre a qual ramo de atividade a empresa pertence (comércio, indústria ou serviços). Teste a hipótese nula padrão em um teste bicaudal de que o endividamento de longo prazo das empresas do setor de comércio é estatisticamente igual ao endividamento do setor de serviços. Em um nível de 10% de significância, a hipótese nula é rejeitada? E em um nível de 5% de significância?
   - **e.** Reporte a diferença na média dos grupos.
   - **f.** O teste t pode ser estimado pressupondo variâncias equivalentes ou variâncias diferentes entre os grupos. Qual dos dois testes é mais adequado para a amostra estudada?

3. Ainda por meio do arquivo **endividamento.dta**, pede-se:
   - **a.** Qual é o número total de observações de cada grupo (ramo de atividade)? Qual dos grupos apresenta a menor e a maior média?
   - **b.** Realize a análise da variância para os dados. Quais são os graus de liberdade para o numerador da estatística F? E do denominador?
   - **c.** Qual é o p-valor para a hipótese nula de que todas as médias são estatisticamente iguais? A hipótese nula é rejeitada a um nível de 10%? E a 2%?

4. O arquivo **tv.dta** contém dados obtidos de uma empresa cujo objetivo consiste em avaliar a preferência do consumidor no momento de aquisição de um aparelho de televisão, com base no preço e na qualidade do suporte técnico, a partir de variáveis referentes a classe social e sexo. Com base nessas informações, elabore e interprete a MANOVA.

CAPÍTULO 4

Regressão Linear

A regressão linear é a técnica que busca estimar o valor esperado para uma variável, denominada dependente, a partir da variação de outra(s) variável(is), denominada(s) explicativa(s), considerando a variável dependente como uma função linear da(s) explicativa(s).

Neste capítulo apresentaremos os principais comandos para a estimação de uma regressão linear, utilizando tanto a regressão simples quanto a regressão múltipla. Abordaremos, também, a análise dos resíduos e a utilização da técnica para a previsão de valores.

Usaremos em nossos exemplos a base de dados **cidades.dta**. A referida base possui 153 observações sobre valores médios simulados sobre o censo de 153 cidades. É composta pelas variáveis descritas no Quadro 4.1.

Quadro 4.1 Variáveis que compõem a base de dados **cidades.dta**

Variável	Descrição	Tipo
mun	Código de identificação do município	Qualitativa
regiao	Região (em total de três regiões)	Qualitativa
medpop	Idade mediana da população	Quantitativa
mat	Taxa de matrimônio (razão do número de matrimônios por 100 mil habitantes)	Quantitativa
div	Taxa de divórcio (razão do número de divórcios por 100 mil habitantes)	Quantitativa

Na janela de comandos do aplicativo Stata®, solicitaremos a abertura da base de dados **cidades.dta**, utilizando o comando **use**.

Na janela de comandos digitaremos o seguinte (lembre-se de informar o endereço completo de localização do arquivo **cidades.dta**):

RESULTADOS 4.1 Abertura do arquivo cidades.dta.

```
. use "cidades.dta"
(Dados simulados sobre munícipios)
```

4.1. REGRESSÃO LINEAR SIMPLES

Na regressão linear simples temos apenas uma variável explicativa. O modelo regressivo simples se assemelha a uma função do primeiro grau, conforme apresentamos no Quadro 4.2.

Quadro 4.2 Modelo de regressão linear simples

$$y = \alpha + \beta x + \varepsilon \quad \text{[Equação 4.1]}$$

Em que:
y: é a variável dependente;
x: é a variável explicativa;
α e β: são os parâmetros da regressão; e
ε: termo de erro da regressão.

No Stata®, para estimar uma regressão linear devemos utilizar o comando **regress** (Sintaxe 4.1).

SINTAXE 4.1 Comando regress.
regress depvar indepvars [, nocons] [, beta] [, level (#)]
Em que:
- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- nocons: Opção a ser utilizada quando não se deseja a presença da constante no modelo regressivo.
- beta: Opção que exibe os coeficientes padronizados.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95%.

O estimador utilizado pelo comando **regress** é o estimador dos mínimos quadrados ordinários que, para uma regressão simples, possui os seguintes pressupostos:

1. A variável dependente deve apresentar distribuição normal.
2. Os resíduos estimados devem possuir distribuição normal.
3. Não deve haver correlação elevada entre os resíduos e a variável explicativa (resíduos homocedásticos).
4. Caso estejamos lidando com uma série temporal (ou seja, as observações variam em função do tempo), os resíduos não poderão ser autocorrelacionados (ausência de autocorrelação dos resíduos).

O poder explicativo de um modelo regressivo é dado pela estatística denominada R^2. O R^2 representa o percentual de variância da variável dependente captado pelas variáveis explicativas. No caso da regressão linear simples, o R^2 representa a correlação simples ao quadrado entre a variável dependente e a explicativa (FÁVERO *et al.*, 2009).

Para verificar a significância conjunta das variáveis explicativas é utilizado o teste F, cuja estatística possui distribuição F com k-1 graus de liberdade (g.l.) no numerador e n-k g.l. no denominador. O número de parâmetros estimados é representado por k, enquanto n compreende o número total de observações. Na regressão linear simples o número de parâmetros será sempre dois. As hipóteses do teste são: H_0: todos os parâmetros β são estatisticamente iguais a zero, e H_1: há pelo menos um parâmetro β estatisticamente diferente de zero.

Existe ainda o teste de significância individual que, na regressão por mínimos quadrados, é o teste t. Este teste é utilizado para verificar se o parâmetro estimado pode ser considerado estatisticamente significativo ou não, em um determinado nível de significância. Na regressão linear simples são realizados dois testes t: um para o intercepto (cujas hipóteses são: $H_0: \alpha = 0$ e $H_1: \alpha \neq 0$) e outro para o coeficiente da variável explicativa (cujas hipóteses são: $H_0: \beta = 0$ e $H_1: \beta \neq 0$).

Na próxima seção passaremos a estimar os parâmetros utilizando uma regressão linear simples.

4.2. ESTIMAÇÃO DOS PARÂMETROS

A partir da base de dados em uso neste capítulo, imaginemos a seguinte situação: O governo está desenvolvendo um estudo sobre o número de divórcios visando capacitar a estrutura judiciária no sentido de prestar melhores serviços à população. Para resolver tal questão, utiliza como variável explicativa a taxa de matrimônios.

Primeiramente, iremos analisar se as variáveis *div* e *mat* estão correlacionadas, a fim de verificarmos a possibilidade de se utilizar a técnica de regressão linear simples. Para tanto, usaremos o seguinte comando:

pwcorr div mat, sig

De acordo com o resultado apresentado (Resultados 4.2), as variáveis estão fortemente correlacionadas, o que é um indicativo de que seja possível estabelecer uma relação linear entre ambas.

Todavia, destacamos que o objetivo desse exemplo é meramente didático e não estamos adentrando em outro uso bastante comum da regressão linear, que é a avaliação da relação de causa e efeito, geralmente embasada em uma teoria subjacente. No exemplo, a regressão está sendo realizada com o objetivo de se estabelecer uma relação linear entre duas variáveis, sem, contudo, descrever uma relação de causa e efeito.

RESULTADOS 4.2 Análise da correlação entre as variáveis *div* e *mat*.

```
. pwcorr div mat, sig

             |      div      mat
-------------+------------------
         div |   1.0000
             |
             |
         mat |   0.9321   1.0000
             |   0.0000
             |
```

Em relação à correlação, podemos solicitar o comando **pwcorr** por meio da barra de menus, selecionando as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Pairwise correlations*. Aparecerá uma janela, conforme a Figura 4.1.

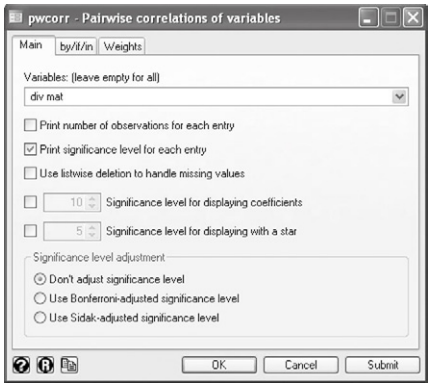


Figura 4.1 ***Janela de configurações do comando*** **pwcorr.**

Para estimarmos a regressão na qual a variável dependente é a taxa de divórcio e a explicativa, a taxa de matrimônio, digitaremos o seguinte comando:

regress div mat

Na primeira parte do resultado da regressão estimada são evidenciados, principalmente, o número de observações, a estatística e o p-valor do teste F e o R^2. Na segunda parte, são exibidos os parâmetros estimados, os erros-padrão de cada parâmetro, as estatísticas e os p-valores do teste t e os intervalos de confiança (Resultados 4.3).

RESULTADOS 4.3 Resultados da regressão linear simples.

```
. regress div mat

      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F(  1,   151) =  999.76
       Model |  883.547715     1  883.547715            Prob > F      =  0.0000
    Residual |  133.447686   151  .883759511           R-squared     =  0.8688
-------------+------------------------------           Adj R-squared =  0.8679
       Total |  1016.9954    152  6.69075921           Root MSE      =  .94008

------------------------------------------------------------------------------
         div |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mat |   .3436585   .0108687    31.62   0.000      .322184    .3651329
       _cons |    1.13232   .6025058     1.88   0.062    -.0581099    2.322751
------------------------------------------------------------------------------
```

Para acessar o comando **regress**, por intermédio da barra de menus, devemos clicar nas seguintes opções: *Statistics* → *Linear models and related* → *Linear regression*. Surgirá uma janela, conforme a Figura 4.2.

Na próxima seção passaremos a analisar os resultados da regressão linear simples.

Figura 4.2 ***Janela de configurações do comando*** **regress.**

4.3. RESULTADOS DA REGRESSÃO LINEAR SIMPLES

Na primeira parte dos Resultados 4.3, notamos que foram utilizadas 153 observações para a estimação dos dois parâmetros. O R^2, de 0,869, equivale ao quadrado da correlação linear entre as duas variáveis ($0{,}9321^2 = 0{,}869$). Em outras palavras, 86,9% da variação do comportamento de *div* pode ser explicado pelo comportamento de *mat*.

Todavia, apenas uma estatística R^2 com um alto valor não é suficiente para atestarmos sobre a significância da regressão estimada. O teste F resultou em uma estatística de 999,76 que, em uma distribuição $F_{1,151}$ (graus de liberdade do numerador: k-1 = 2-1 = 1; graus de liberdade do denominador: n-k = 153-2 = 151), retorna um p-valor inferior a 0,001.

Tal resultado nos leva à rejeição da hipótese nula de que todos os parâmetros sejam estatisticamente iguais a zero, o que, no caso da regressão linear simples, representa que o coeficiente da variável explicativa possui significância estatística.

Em relação ao teste t, verificamos que o coeficiente da variável explicativa é considerado estatisticamente significativo, pois, com um p-valor inferior a 0,001, rejeita-se a hipótese de que esse parâmetro seja igual a zero, diferentemente do que acontece com o intercepto, cujo p-valor é de 0,062. Uma propriedade em relação à regressão linear simples é que a estatística t do coeficiente da variável explicativa ao quadrado é igual à estatística F [$(31,62)^2$ = 999,76].

Em todas as análises realizadas, utilizamos o nível de significância de 5%.

De acordo com o modelo estimado, a cada alteração em uma unidade na taxa de matrimônio ocorre 0,343 de variação na taxa de divórcio.

4.4. VALORES PREVISTOS E RESÍDUOS

Antes de fazermos qualquer inferência com os resultados de uma regressão, seja ela simples ou múltipla, precisamos nos certificar de que os pressupostos da técnica são atendidos.

Dessa forma, precisamos realizar alguns testes. No nosso exemplo, verificaremos se os resíduos possuem distribuição normal e se são homocedásticos. Inicialmente, utilizaremos o comando **predict** para a geração da série de resíduos (Sintaxe 4.2).

SINTAXE 4.2 Comando predict.

predict newvar [, residual] [, rstandard] [, xb]

Em que:

- newvar: Nome da nova variável que armazenará os valores previstos.
- residual: Opção a ser utilizada para a geração dos resíduos da regressão.
- rstandard: Opção a ser utilizada para a geração dos resíduos padronizados da regressão.
- xb: Opção a ser utilizada para a geração dos valores estimados da variável dependente.

Na janela de comandos do Stata®, informaremos o seguinte:

predict resid, residual

RESULTADOS 4.4 Execução do comando predict com a opção residual.

```
. predict resid, residual
```

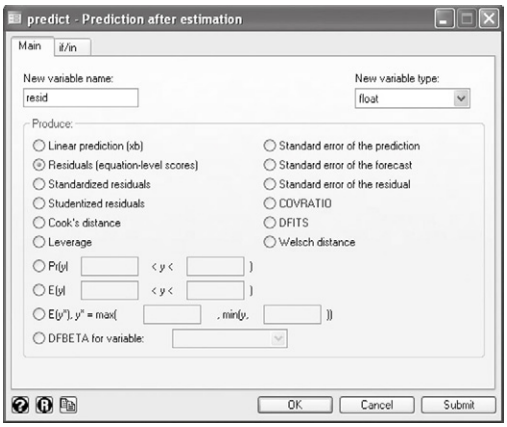


Figura 4.3 ***Janela de configurações do comando*** **predict.**

Esse comando pode ser acessado via barra de menus, por meio dos seguintes passos: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Irá surgir uma janela, conforme a Figura 4.3.

Após gerarmos a variável *resid*, que contém os resíduos da regressão, iremos solicitar o teste Shapiro-Francia para verificar se a mesma possui uma distribuição normal. Utilizaremos o seguinte comando:

sfrancia resid

RESULTADOS 4.5 Resultado do teste de normalidade para a variável *resid*.

```
. sfrancia resid

                  Shapiro-Francia W' test for normal data

    Variable |     Obs        W'           V'         z        Prob>z
-------------+--------------------------------------------------------
       resid |     153     0.98886       1.445      0.748     0.22730
```

De acordo com o resultado do teste Shapiro-Francia, verificamos, com probabilidade de 0,22, que os resíduos possuem uma distribuição normal, não havendo rejeição da hipótese nula (Resultados 4.5).

A variável dependente *div* também apresenta distribuição normal, com probabilidade de 0,28. O resultado do teste não será aqui apresentado, mas o pesquisador pode obtê-lo por meio da aplicação do comando **sfrancia div**.

Apenas para relembrar, o teste Shapiro-Francia pode ser acessado mediante a seleção das seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Shapiro-Francia normality test*. Poderiam ter sido usados outros testes de normalidade, à escolha do pesquisador, conforme vimos no Capítulo 2.

Para verificarmos a homocedasticidade dos resíduos, ou seja, se os mesmos possuem variância constante, utilizaremos o teste Breusch-Pagan por meio do comando **estat hettest** (Sintaxe 4.3) (nas versões mais antigas do Stata®, apenas **hettest**).

SINTAXE 4.3 Comando estat hettest.

estat hettest [varlist]

Em que:

- varlist: Lista contendo as variáveis explicativas que serão utilizadas no cômputo da estatística do teste. Caso não seja informada nenhuma variável, o Stata® utilizará as variáveis explicativas da última regressão estimada.

O teste Breusch-Pagan possui as seguintes hipóteses: H_0: os resíduos são homocedásticos, e H_1: os resíduos são heterocedásticos. Informaremos, na janela de comandos, o seguinte:

estat hettest

Com um p-valor superior a 0,17, verificamos que a hipótese nula do teste Breusch-Pagan não foi rejeitada e, assim sendo, os resíduos são considerados homocedásticos (Resultados 4.6).

RESULTADOS 4.6 Resultado do teste Breusch-Pagan.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of div

         chi2(1)      =     1.85
         Prob > chi2  =   0.1733
```

Sendo a variável dependente normal e os resíduos normais e homocedásticos, verificamos que todos os pressupostos do estimador dos mínimos quadrados foram respeitados para a regressão linear simples e, portanto, os resultados estimados são válidos e possíveis para utilização em inferências.

Para executarmos o comando **estat hettest**, utilizando a barra de menus, devemos clicar nas seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Aparecerá uma janela, conforme a Figura 4.4.

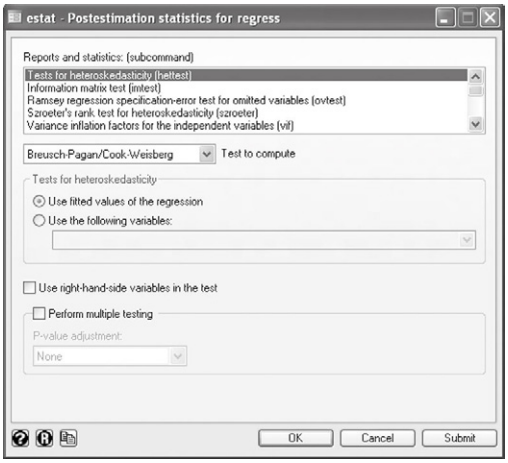


Figura 4.4 ***Janela de configurações do comando*** **estat** ***selecionando-se a opção*** **hettest.**

Os valores estimados para a variável dependente ($\hat{Y}$) podem ser obtidos, inclusive para cada observação da amostra, por meio do comando **predict**, conforme demonstrado a seguir:

predict estimat, xb

RESULTADOS 4.7 Execução do comando predict com a opção xb.

```
. predict estimat, xb
```

O Stata® gerará uma série de observações, utilizando os parâmetros da última regressão estimada. Mais adiante, faremos uso dos valores estimados para a variável dependente.

Para acessar o comando anteriormente executado, por intermédio da barra de menus, será necessário acessar as seguintes opções: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Será exibida uma janela, conforme a Figura 4.5.

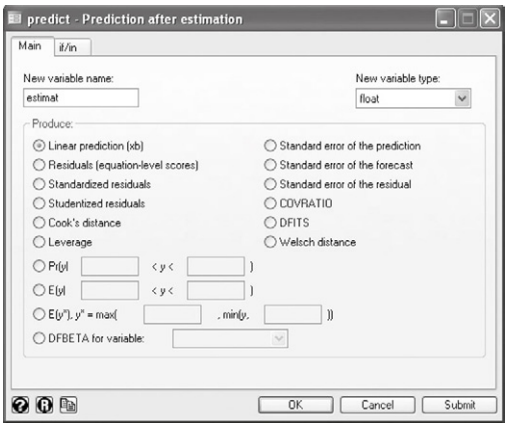


Figura 4.5 ***Janela de configurações do comando*** **predict.**

4.5. GRÁFICOS E TABELAS

Podem ser utilizados os gráficos que mostram a relação entre as duas variáveis para melhor entender os procedimentos realizados durante a estimação da regressão.

O primeiro gráfico que iremos analisar trata-se do gráfico de dispersão utilizando as variáveis dependente e explicativa. Conforme vimos no Capítulo 2, um gráfico de dispersão pode ser gerado utilizando-se o comando **twoway scatter**.

Continuando com o nosso exemplo de regressão linear simples, iremos gerar o gráfico de dispersão entre as variáveis *div* e *mat* (Figura 4.6). Para tanto, informaremos na janela de comandos do Stata® o seguinte:

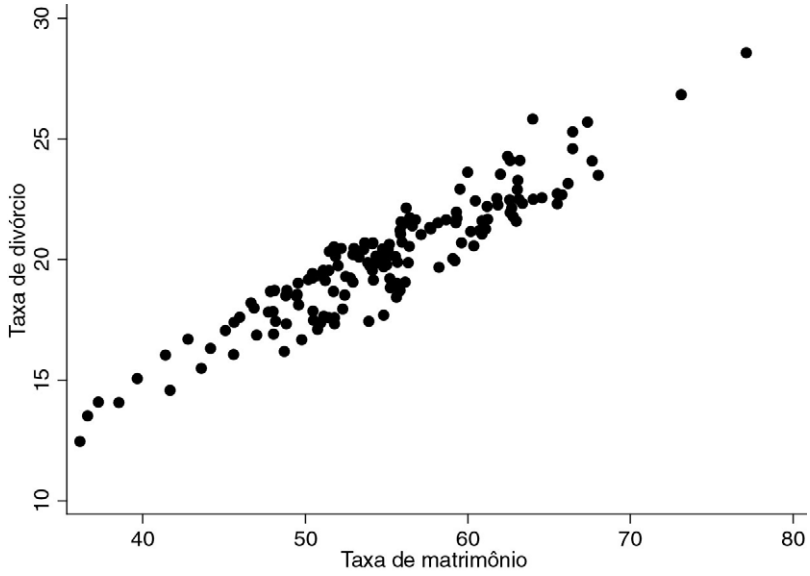


Figura 4.6 ***Gráfico de dispersão entre as variáveis*** **div** ***e*** **mat.**

twoway scatter div mat

RESULTADOS 4.8 Gerando o gráfico de dispersão.

```
. twoway scatter div mat
```

Observamos no gráfico de dispersão alguns pontos mais isolados. Iremos combinar o gráfico de dispersão com o gráfico de linha, para verificarmos visualmente o resultado da regressão estimada (Figura 4.7). Usaremos, novamente, o comando **twoway** combinando os gráficos **scatter** e **line**. Devemos digitar o seguinte comando:

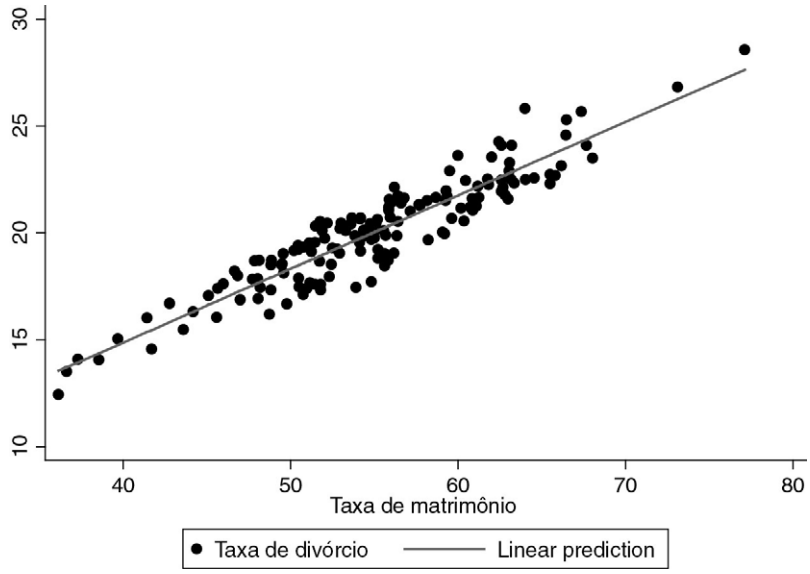


Figura 4.7 ***Gráfico de dispersão entre as variáveis*** **div** ***e*** **mat** ***e de linha entre as variáveis*** **estimat** ***e*** **mat.**

twoway (scatter div mat) (line estimat mat, sort)

RESULTADOS 4.9 Gerando o gráfico de dispersão e de linha.

```
. twoway (scatter div mat) (line estimat mat, sort)
```

A partir da análise gráfica entre a dispersão das variáveis observadas e a reta estimada da regressão, verificamos a presença de alguns pontos dispersos.

Caso desejássemos gerar o gráfico, a partir da barra de menus, deveríamos selecionar as seguintes opções: *Graphics* → *Twoway graph (scatter, line, etc.)*. Será exibida uma janela, conforme as Figuras 4.8 (*scatter*) e 4.9 (*line*).

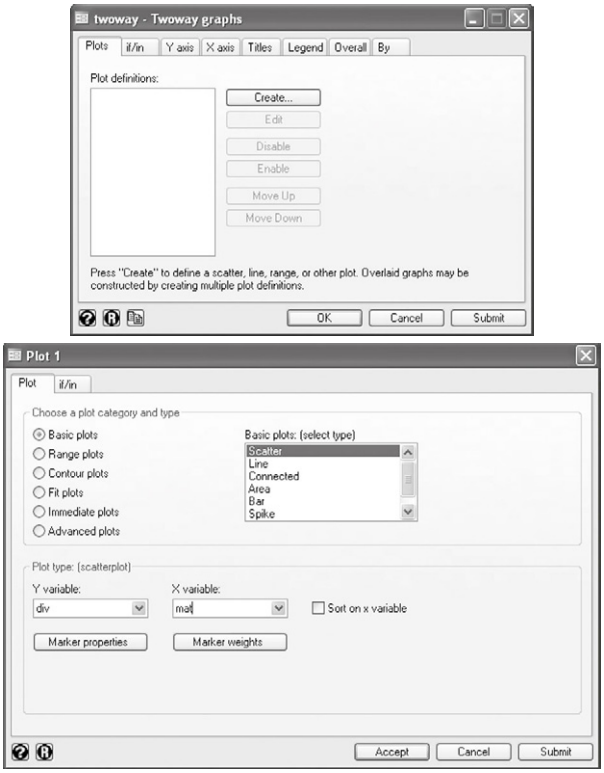


Figura 4.8 ***Janela de configurações do comando* twoway *– adição do primeiro gráfico (*scatter*).***

4.6. REGRESSÃO MÚLTIPLA

Na regressão linear múltipla temos mais de uma variável explicativa (Quadro 4.3).

Para estimar uma regressão linear múltipla no Stata® devemos, também, utilizar o comando **regress**.

O estimador utilizado pelo comando **regress** é o estimador dos mínimos quadrados ordinários. No caso de uma regressão linear múltipla, esse estimador possui os seguintes pressupostos:

1. A variável dependente deve apresentar distribuição normal.
2. Os resíduos estimados devem possuir distribuição normal.
3. Não devem haver correlações elevadas entre os resíduos e cada uma das variáveis explicativas (resíduos homocedásticos).

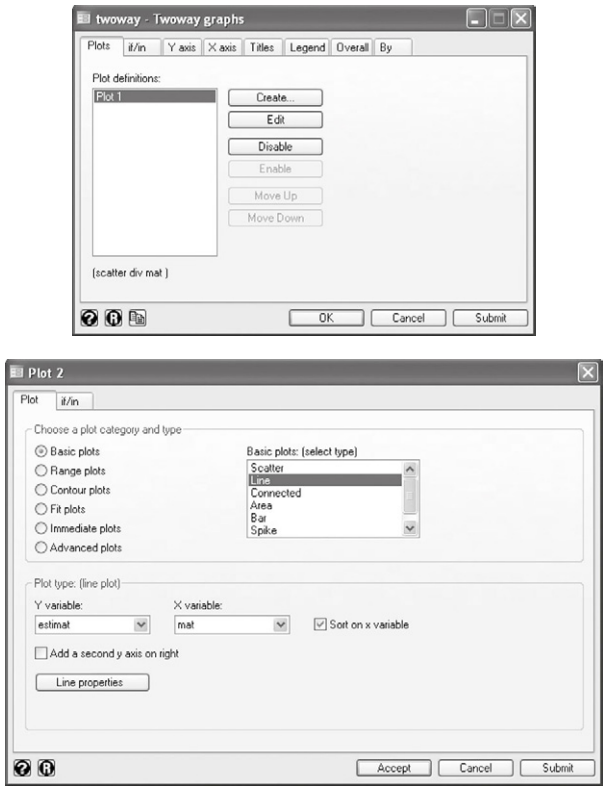


Figura 4.9 *Janela de configurações do comando* **twoway** *– adição do segundo gráfico (*line*).*

Quadro 4.3 Modelo de regressão linear múltipla

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \text{[Equação 4.2]}$$

Em que:
y: é a variável dependente;
x_i: são as variáveis explicativas, com $i = 1, 2, \ldots, k$;
α e β_i: são os parâmetros da regressão; e
ε: termo de erro da regressão.

4. Não deve haver correlações elevadas entre as variáveis explicativas (ausência de multicolinearidade).
5. Caso estejamos lidando com uma série temporal (ou seja, as observações variam em função do tempo), os resíduos não poderão ser autocorrelacionados (ausência de autocorrelação dos resíduos).

O poder explicativo de um modelo regressivo é dado pela estatística R^2. Porém, na regressão linear múltipla pode também ser definido o R^2 Ajustado. Como o R^2 sempre aumentará, mesmo que minimamente, quando uma nova variável é adicionada

ao modelo, o que ocorrerá quando deixarmos de utilizar a regressão linear simples para fazer uso da regressão múltipla, deve-se ponderar o seu cálculo pelo número de graus de liberdade do modelo, a fim de que o mesmo possa ser comparado com modelos com diferentes graus de liberdade. Esta ponderação é feita no cálculo do R^2 Ajustado.

Para verificarmos a significância conjunta das variáveis explicativas, é utilizado o teste F, cujas hipóteses são: H_0: todos os parâmetros β são estatisticamente iguais a zero, e H_1: há pelo menos um parâmetro β estatisticamente diferente de zero. O teste t é o teste de significância individual. Na regressão linear múltipla são realizados os testes t considerando as seguintes hipóteses: (i) para o intercepto: $H_0: \alpha = 0$ e $H_1: \alpha \neq 0$; e (ii) para os coeficientes das variáveis explicativas: $H_0: \beta_i = 0$ e $H_1: \beta_i \neq 0$.

Voltando ao nosso exemplo referente à regressão linear simples e utilizando outras variáveis contidas na base de dados, vamos passar para o modelo de regressão múltipla.

Duas variáveis ainda não utilizadas nos chamam a atenção: *medpop* e *region*. A primeira compreende a mediana da idade da população de cada município, sendo, portanto, quantitativa, e a segunda trata da região onde o município está localizado, sendo uma variável categórica (qualitativa).

Primeiramente, iremos analisar se as variáveis *div*, *mat* e *medpop* estão correlacionadas para, então, verificarmos a possibilidade de utilizar a técnica de regressão linear. Lembremos que é importante que as variáveis explicativas estejam correlacionadas com a dependente, mas não fortemente correlacionadas entre si. Para tanto, usaremos o seguinte comando:

pwcorr div mat medpop, sig

RESULTADOS 4.10 Análise da correlação entre as variáveis *div*, *mat* e *medpop*.

```
. pwcorr div mat medpop, sig

             |      div      mat   medpop
-------------+---------------------------
         div |   1.0000
             |
             |
         mat |   0.9321   1.0000
             |   0.0000
             |
      medpop |   0.0911   0.0855   1.0000
             |   0.2627   0.2932
             |
```

Segundo as correlações e os níveis de significância apresentados nos Resultados 4.10, verificamos que: (i) não há correlação significativa entre as duas variáveis explicativas, o que não geraria problemas de multicolinearidade; e (ii) todavia, não há correlação significativa, também, entre a variável *medpop* e a variável dependente, demonstrando não haver uma relação linear entre essas variáveis.

Mesmo diante da ausência de relação linear entre as variáveis *div* e *medpop*, para fins didáticos iremos incluir a última variável no modelo de regressão simples, transformando-o em uma regressão múltipla.

Iremos adicionar na regressão a variável *regiao*, que é uma variável qualitativa. Como não se pode adicionar uma variável categórica diretamente em uma regressão, pois todas as variáveis explicativas precisam ser métricas, podemos utilizar variáveis *dummies* oriundas da variável categórica original. No Stata® podemos utilizar o prefixo **i.** para que sejam inseridas automaticamente variáveis *dummies* criadas a partir de uma variável categórica.

Agora, iremos digitar o seguinte comando:

regress div mat medpop i.regiao

RESULTADOS 4.11 Resultados da regressão linear múltipla.

```
. regress div mat medpop i.regiao

      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F(  4,   148) = 1727.78
       Model |  995.673292     4  248.918323            Prob > F      =  0.0000
    Residual |  21.3221086   148  .144068301           R-squared     =  0.9790
-------------+------------------------------           Adj R-squared =  0.9785
       Total |  1016.9954    152  6.69075921           Root MSE      =  .37956

------------------------------------------------------------------------------
         div |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mat |   .3880443   .0046881    82.77   0.000     .3787801    .3973085
      medpop |   .0061801   .0086095     0.72   0.474    -.0108332    .0231934
             |
      regiao |
          2  |   1.002185   .0756387    13.25   0.000     .8527135    1.151656
          3  |  -1.196083   .0791239   -15.12   0.000    -1.352442   -1.039725
             |
       _cons |  -1.720251   .6058454    -2.84   0.005    -2.917476   -.5230266
------------------------------------------------------------------------------
```

Antes de analisarmos os resultados estimados pela regressão, vamos realizar os testes necessários para verificar se os pressupostos da regressão foram respeitados.

Começaremos com os testes para detecção da normalidade e da homocedasticidade dos resíduos. Utilizaremos os seguintes comandos:

predict res1, residual

sfrancia res1

estat hettest

RESULTADOS 4.12 Testes acessórios para a regressão linear múltipla.

```
. predict res1, residual

. sfrancia res1

                  Shapiro-Francia W' test for normal data

    Variable |     Obs        W'           V'         z        Prob>z
-------------+--------------------------------------------------------
        res1 |     153    0.98951       1.361      0.626     0.26558

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of div

         chi2(1)      =     0.21
         Prob > chi2  =   0.6469
```

Os resultados dos testes Shapiro-Francia (normalidade) e Breusch-Pagan (heterocedasticidade) indicam que os resíduos são normais e homocedásticos, sendo as respectivas hipóteses nulas não rejeitadas (Resultados 4.12).

Além da normalidade e da homocedasticidade dos resíduos, na regressão linear múltipla precisamos verificar se há problemas de multicolinearidade, ou seja, as variáveis explicativas possuem fortes correlações entre si. Não há um teste específico, porém, uma estatística bastante utilizada é o VIF (*variance inflation factor* ou fator de inflação da variância). Segundo Gujarati (2011), um VIF acima de 10 é indicativo de multicolinearidade. Fávero *et al.* (2009), ao serem até mais rigorosos, já argumentam que um VIF acima de 5 também pode causar problemas de multicolinearidade.

No Stata® podemos visualizar a estatística VIF por intermédio do comando **estat vif** (Sintaxe 4.4) (nas versões mais antigas, apenas **vif**).

SINTAXE 4.4 Comando estat vif.

estat vif [, uncentered]

Em que:

- uncentered: Opção que poderá ser utilizada quando for omitida a constante da regressão.

Para verificar se há problemas de multicolinearidade, iremos solicitar as estatísticas VIF por meio do seguinte comando:

estat vif

RESULTADOS 4.13 Estatísticas VIF.

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
         mat |      1.14    0.876203
      medpop |      1.01    0.992610
      regiao |
           2 |      1.43    0.699231
           3 |      1.43    0.698576
-------------+----------------------
    Mean VIF |      1.25
```

Observamos que todas as estatísticas VIF foram inferiores a 5 (Resultados 4.13). Assim, concluímos que não há problemas de multicolinearidade e podemos passar à análise dos resultados da regressão múltipla (Resultados 4.11).

O R^2, de 0,979, é superior ao da regressão linear simples, conforme já discutido.

O teste F resultou em uma estatística de 1.727,78 que, em uma distribuição $F_{4,148}$ (graus de liberdade do numerador: k-1 = 5-1 = 4; graus de liberdade do denominador: n-k = 153-5 = 148), retornou um p-valor inferior a 0,001. Esse resultado nos leva à rejeição da hipótese nula de que todos os parâmetros sejam estatisticamente iguais a zero, ou seja, de que existe pelo menos um coeficiente das variáveis explicativas que é estatisticamente significante a 5%.

Em relação ao teste t, verificamos que o coeficiente da variável *medpop* não se mostrou estatisticamente significante a 5% (0,05), pois apresentou p-valor superior a 0,47. Os coeficientes das variáveis *dummies* associadas às categorias 2 e 3 da variável *regiao* mostraram-se estatisticamente significantes a 5%.

Em relação à variável *medpop*, confirmamos aquilo que havíamos discutido quando analisamos a correlação entre essa variável e a dependente. Em relação às variáveis *dummies*, os resultados dos testes t nos levam à conclusão, considerando a categoria 1 da variável *regiao* como grupo de referência, de que há diferenças dessa região em relação às demais para o comportamento da variável *div*.

Em função dos resultados obtidos, iremos retirar a variável explicativa *medpop* e efetuaremos nova estimação (Resultados 4.14), digitando o seguinte comando:

regress div mat i.regiao

RESULTADOS 4.14 Resultados da regressão linear múltipla.

```
. regress div mat i.regiao

      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F(  3,   149) = 2311.05
       Model |  995.599057     3  331.866352            Prob > F      =  0.0000
    Residual |  21.3963435   149  .143599621            R-squared     =  0.9790
-------------+------------------------------           Adj R-squared =  0.9785
       Total |   1016.9954   152  6.69075921            Root MSE      =  .37895

------------------------------------------------------------------------------
         div |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mat |   .3883247   .0046642    83.26   0.000     .3791083    .3975411
             |
      regiao |
           2 |   1.002376   .0755151    13.27   0.000     .8531568    1.151594
           3 |  -1.196387    .078994   -15.15   0.000    -1.352481   -1.040294
             |
       _cons |  -1.328611   .2629691    -5.05   0.000    -1.848241   -.8089803
------------------------------------------------------------------------------
```

Quando queremos trabalhar com apenas algumas categorias de uma variável ou desejamos criar variáveis *dummies* para testar seus efeitos isoladamente, no Stata® o fazemos por meio do comando **xi** (Sintaxe 4.5).

SINTAXE 4.5 Comando xi.

xi i.varname

Em que:

- varname: Nome da variável categórica que será convertida em variáveis *dummies*.

Vamos, inicialmente, solicitar a criação das variáveis *dummies*, visto que utilizaremos apenas a categoria relativa à região 2 na regressão múltipla. Na janela de comandos devemos digitar o seguinte:

xi i.regiao

RESULTADOS 4.15 Criando variáveis *dummies* a partir de uma variável categórica.

```
. xi i.regiao
i.regiao          _Iregiao_1-3        (naturally coded; _Iregiao_1 omitted)
```

Podemos notar que foram criadas duas variáveis *dummies*, com os nomes de *_Iregiao_2* e *_Iregiao_3*. A primeira categoria da variável *regiao* é considerada a referência. Vamos para a estimação da regressão, digitando o seguinte comando:

regress div mat _Iregiao_2

RESULTADOS 4.16 Resultados da regressão linear múltipla.

```
. regress div mat _Iregiao_2

      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F(  2,   150) = 1328.78
       Model |  962.660176     2  481.330088            Prob > F      =  0.0000
    Residual |  54.3352244   150  .362234829            R-squared     =  0.9466
-------------+------------------------------           Adj R-squared =  0.9459
       Total |   1016.9954   152  6.69075921            Root MSE      =  .60186

------------------------------------------------------------------------------
         div |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mat |   .3758989   .0072923    51.55   0.000     .3614899    .3903078
  _Iregiao_2 |   1.553288   .1051053    14.78   0.000     1.345609    1.760966
       _cons |  -1.229469   .4175309    -2.94   0.004     -2.05447   -.4044671
------------------------------------------------------------------------------
```

Para analisarmos os resultados da regressão, precisarmos nos certificar de que os pressupostos foram atendidos e, portanto, solicitaremos testes e estatísticas por meio dos seguintes comandos:

predict res2, residual
sfrancia res2
estat hettest
estat vif

RESULTADOS 4.17 Testes e estatísticas acessórios para a regressão linear múltipla.

```
. predict res2, residual

. sfrancia res2

                  Shapiro-Francia W' test for normal data

    Variable |    Obs        W'           V'         z        Prob>z
-------------+-------------------------------------------------------
        res2 |    153    0.99448       0.716    -0.679     0.75147

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of div

         chi2(1)      =     1.97
         Prob > chi2  =   0.1609

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
  _Iregiao_2 |      1.10    0.910502
         mat |      1.10    0.910502
-------------+----------------------
    Mean VIF |      1.10
```

O teste Shapiro-Francia resultou em um p-valor superior a 0,75, o que implica a não rejeição da hipótese nula de que os resíduos possuem distribuição normal. O teste Breusch-Pagan resultou em um p-valor superior a 0,16, indicando que os resíduos são homocedásticos. As estatísticas VIF foram inferiores a 5, descartando-se o problema da multicolinearidade. Verificamos, então, que os pressupostos foram respeitados e passaremos às análises das estimações realizadas (Resultados 4.17).

Os coeficientes R^2 e R^2 Ajustado foram similares aos obtidos na primeira estimação com uma regressão múltipla que fizemos com todas as variáveis explicativas. Podemos notar que a ausência da variável *medpop* e da *dummy* relativa à categoria 3 da variável *regiao* não afetaram consideravelmente o poder explicativo do atual modelo (Resultados 4.16).

O teste F resultou em um p-valor inferior a 0,001, implicando a rejeição da hipótese nula de que todos os coeficientes estimados das variáveis explicativas sejam estatisticamente iguais a zero. Individualmente, por intermédio do teste t, verificamos que todas as variáveis explicativas e a constante foram consideradas significativas a um nível de 5% (Resultados 4.16). Assim, o modelo regressivo estimado pode ser representado pela seguinte equação:

$$\text{estimação de } div = -1,229 + 0,375.mat + 1,553.\_\text{Iregiao}\_2 \quad \text{[Equação 4.3]}$$

Segundo o modelo estimado, a cada alteração em uma unidade na taxa de matrimônio ocorre 0,375 de variação na taxa de divórcio, mantidas as demais condições constantes.

Todavia, verificamos que, se um município estiver situado na região 2, a sua taxa de divórcio será alterada em relação aos municípios situados nas regiões 1 e 3. Se um município estiver situado na região 2, a taxa de divórcio sofrerá uma variação de 1,553. Em outras palavras, havendo dois municípios com a mesma taxa de matrimônio, porém, um localizado nas regiões 1 ou 3 e o outro localizado na região 2, esse último terá uma taxa de divórcio superior em 1,553 unidades em relação ao primeiro.

Passamos agora a apresentar como realizar os procedimentos anteriores por meio da barra de menus. Em relação ao comando **xi**, podemos acessá-lo clicando nas seguintes opções: *Data* → *Create or change data* → *Other variable-creation commands* → *Interaction expansion*. Aparecerá uma janela, conforme a Figura 4.10.

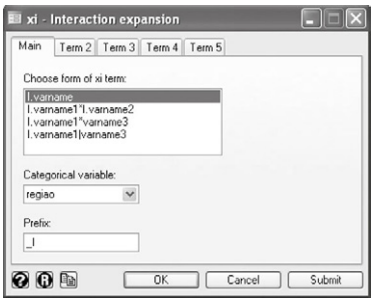


Figura 4.10 ***Janela de configurações do comando*** **xi.**

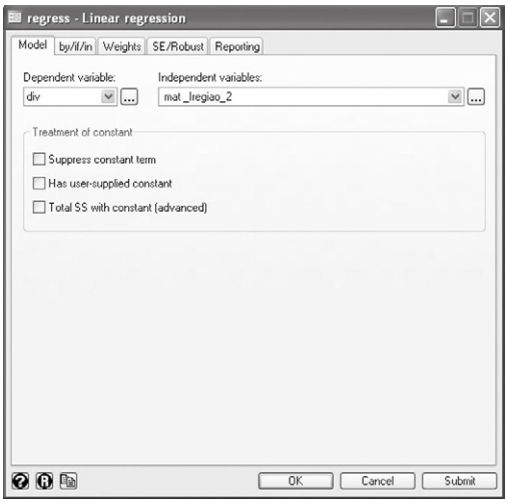


Figura 4.11 ***Janela de configurações do comando*** **regress.**

Em relação ao comando **regress**, precisaremos acessar as seguintes opções: *Statistics* → *Linear models and related* → *Linear regression*. Surgirá uma janela, conforme a Figura 4.11.

Para gerar a série de resíduos da regressão, acessamos o comando **predict**, a partir das seguintes opções: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Irá surgir a janela da Figura 4.12.

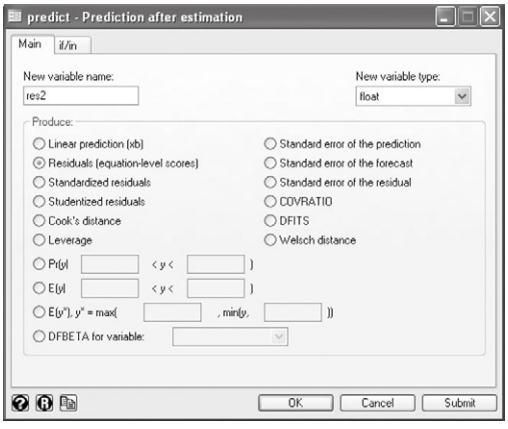


Figura 4.12 ***Janela de configurações do comando*** **predict.**

Para realizar o teste Shapiro-Francia, basta selecionarmos as seguintes opções: *Statistics* → *Summaries, tables, and tests* → *Distributional plots and tests* → *Shapiro-Francia normality test*. Irá surgir a janela da Figura 4.13.

Para realizar o teste Breusch-Pagan, basta selecionarmos as seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Aparecerá a janela da Figura 4.14.

Para obtermos as estatísticas VIF, basta selecionarmos as seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Surgirá a janela da Figura 4.15.

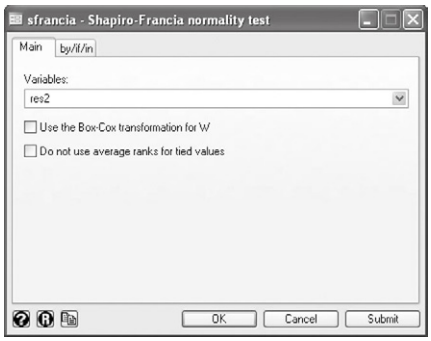


Figura 4.13 *Janela de configurações do comando* **sfrancia.**

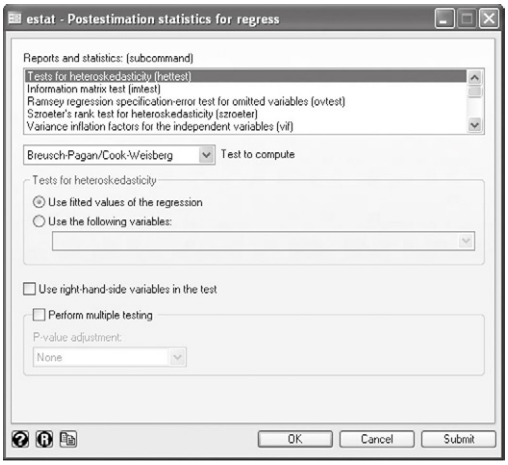


Figura 4.14 *Janela de configurações do comando* **estat** *selecionando-se a opção* **hettest.**

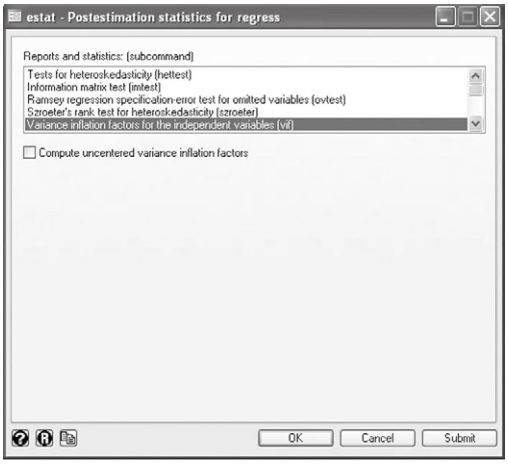


Figura 4.15 *Janela de configurações do comando* **estat** *selecionando-se a opção* **vif.**

4.7. OPÇÕES NA REGRESSÃO LINEAR SIMPLES

Nesta seção, iremos abordar algumas opções disponíveis no comando **regress** utilizando o nosso modelo de regressão linear simples. Entretanto, essas opções são igualmente válidas quando estivermos utilizando uma regressão linear múltipla.

Por padrão, o Stata® utiliza o nível de confiança de 95% para estimar um intervalo de confiança para os parâmetros da regressão. Todavia, caso queiramos trabalhar com outro nível de confiança de, por exemplo, 90%, precisamos utilizar a opção **level**, no comando **regress**.

Voltando ao nosso exemplo de regressão linear simples, modificaremos o nosso nível de significância para 10% (nível de confiança de 90%), utilizando o seguinte comando:

regress div mat, level(90)

RESULTADOS 4.18 Resultados da regressão linear simples.

```
. regress div mat, level(90)

      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F(  1,   151) =  999.76
       Model |  883.547715     1  883.547715           Prob > F      =  0.0000
    Residual |  133.447686   151  .883759511           R-squared     =  0.8688
-------------+------------------------------           Adj R-squared =  0.8679
       Total |  1016.9954    152  6.69075921           Root MSE      =  .94008

------------------------------------------------------------------------------
         div |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
         mat |   .3436585   .0108687    31.62   0.000     .3256706    .3616463
       _cons |    1.13232   .6025058     1.88   0.062     .1351689    2.129472
------------------------------------------------------------------------------
```

Quando comparamos os resultados dessa nova regressão com os obtidos na seção 4.2, verificamos que apenas houve mudança no intervalo de confiança dos parâmetros (Resultados 4.18).

O intervalo de confiança dos parâmetros pode ser utilizado para estimar o intervalo de confiança para a previsão da variável dependente. Assim poderemos definir as equações para os limites inferiores e superiores da regressão, utilizando o intervalo de confiança dos parâmetros.

No nosso exemplo teríamos:

Equação para o limite inferior do valor previsto para a variável dependente:

$$\text{estimação de } div = 0{,}135 + 0{,}326.mat \qquad \text{[Equação 4.4]}$$

Equação para o limite superior do valor previsto para a variável dependente:

$$\text{estimação de } div = 2{,}129 + 0{,}362.mat \qquad \text{[Equação 4.5]}$$

Por exemplo, considerando um nível de confiança de 90%, o intervalo de confiança para a previsão da taxa de divórcio para um município cuja taxa de matrimônio é de 100 seria:

Limite inferior:

$$\text{estimação de } div = 0,135 + 0,326.(100) = 32,735 \quad \text{[Equação 4.6]}$$

Limite superior:

$$\text{estimação de } div = 2,129 + 0,362.(100) = 38,329 \quad \text{[Equação 4.7]}$$

Assim, de acordo com o modelo e considerando um nível de significância de 10%, o real valor de *div* para tal município estaria situado no seguinte intervalo: [32,735; 38,329].

A outra opção se refere à realização do teste Breusch-Godfrey para a detecção de autocorrelação serial, quando utilizamos séries temporais com o comando **regress**.

A base de dados que estamos utilizando neste capítulo compreende uma série transversal, conhecida por *cross-section* (ou seja, somente as observações ou indivíduos analisados variam; o tempo não varia). Todavia, para fins didáticos, iremos transformá-la em uma série longitudinal (ou seja, o tempo passa a variar, mas não os indivíduos ou observações), para a realização do teste para a detecção de autocorrelação dos resíduos.

Criaremos uma variável temporal utilizando o comando **gen**. Informaremos ao Stata®, na janela de comandos, o seguinte:

gen mes = m(2009m12) + _n

RESULTADOS 4.19 Criação de uma variável temporal.

```
. gen mes = m(2009m12) + _n
```

Será criada a variável *mes*, que será utilizada para definir a série como sendo temporal. Para isso, precisaremos do comando **tsset** (Sintaxe 4.6).

SINTAXE 4.6 Comando tsset.

tsset timevar [, options]

Em que:

- timevar: Nome da variável temporal.
- options: Especifica o formato da variável de acordo com a frequência: (i) **daily**: diário; (ii) **weekly**: semanal, (iii) **monthly**: mensal; (iv) **quartely**: quadrimestral; (v) **halfyearly**: semestral; e (vi) **yearly**: anual.

Assim, digitaremos o seguinte comando:

tsset mes, monthly

RESULTADOS 4.20 Definida a série como sendo temporal.

```
. tsset mes, monthly
        time variable:  mes, 2010m1 to 2022m9
                delta:  1 month
```

Para realizar o teste Breusch-Godfrey, utilizaremos o comando **estat bgodfrey** (Sintaxe 4.7) (nas versões mais antigas do Stata®, apenas **bgodfrey**):

SINTAXE 4.7 Comando estat bgodfrey.

estat bgodfrey [, lags(laglist)]

Em que:

- lags: Especifica o número de defasagens (*lags*) que serão testadas para a detecção da autocorrelação. Pode ser informada uma lista de defasagens no lugar do termo **lagslist**.

No nosso exemplo, iremos verificar se existem problemas de autocorrelação serial utilizando até três defasagens. Devemos informar o seguinte comando:

estat bgodfrey, lags (1 2 3)

RESULTADOS 4.21 Teste Breusch-Godfrey.

```
. estat bgodfrey, lags (1 2 3)

Breusch-Godfrey LM test for autocorrelation
-------------------------------------------------------------------------
    lags(p) |          chi2               df                 Prob > chi2
------------+------------------------------------------------------------
          1 |          0.774               1                      0.3790
          2 |          1.042               2                      0.5941
          3 |          1.045               3                      0.7904
-------------------------------------------------------------------------
                    H0: no serial correlation
```

O teste Breusch-Godfrey apresenta a hipótese nula de que os resíduos não são autocorrelacionados na ordem especificada pelo número de defasagens. Caso a base de dados utilizada fosse uma série temporal, e considerando um nível de significância de 5%, verificaríamos que a mesma não apresentaria problemas de autocorrelação serial (Resultados 4.21).

Caso desejássemos utilizar a barra de menus para a seleção dos comandos anteriores, precisaríamos proceder da forma relatada a seguir. Em relação ao comando **gen**, devemos clicar nas seguintes opções: *Data* → *Create or change data* → *Create new variable*. Surgirá a janela da Figura 4.16.

Em relação ao comando **tsset**, o mesmo pode ser acessado por meio da seleção das seguintes opções: *Statistics* → *Time series* → *Setup and utilities* → *Declare dataset to be time-series data*. Aparecerá a janela da Figura 4.17.

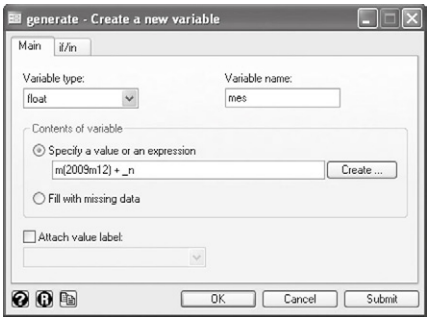


Figura 4.16 ***Janela de configurações do comando*** **gen.**

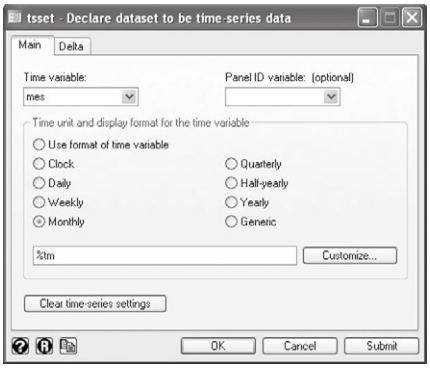


Figura 4.17 ***Janela de configurações do comando*** **tsset.**

Para realizar o teste Breusch-Godfrey, basta selecionarmos as seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Aparecerá a janela da Figura 4.18.

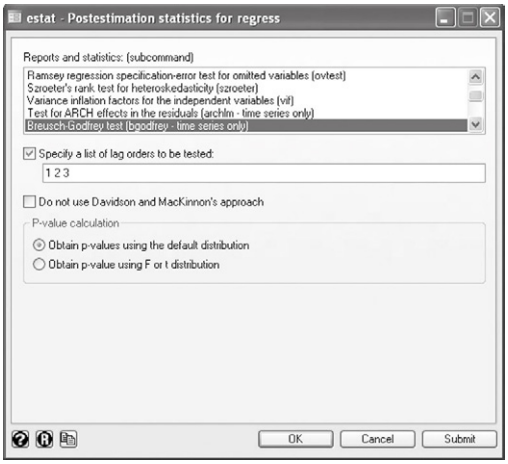


Figura 4.18 ***Janela de configurações do comando*** **estat** ***selecionando-se a opção*** **bgodfrey.**

4.8. EXERCÍCIOS

1. Buscando detectar um padrão de comportamento dos retornos das ações negociadas na bolsa de valores BM&FBovespa, um analista coletou dados referentes aos retornos anuais de 112 empresas (arquivo **retorno.dta**), bem como informações de variáveis julgadas boas preditoras para a variação nos preços das ações, quais sejam:

Variável	Descrição
Tamanho	Tamanho da companhia (logaritmo natural do ativo total da empresa)
Book/Mkt	Quociente entre o valor de livro (Patrimônio Líquido) e o valor de mercado
Beta	Indicador utilizado para calcular o risco das ações
ROA	Retorno sobre Ativo
INV	Crescimento do Ativo Imobilizado entre t e t-1

Com base nessas informações, pede-se:

- **a.** Estime uma regressão em que o retorno é a variável dependente e as demais variáveis apresentadas são as variáveis explicativas. Qual é o número de observações?
- **b.** O erro-padrão serve como uma medida da variabilidade típica do coeficiente de regressão. Quais os erros-padrão das variáveis explicativas da regressão?
- **c.** Qual é o coeficiente de determinação?
- **d.** Qual é o p-valor geral do teste F? Considerando-se 95% de nível de confiança, você rejeita a hipótese nula desta estatística? Qual interpretação pode ser dada diante do resultado do teste?
- **e.** Considerando-se 95% de nível de confiança, você rejeita a hipótese nula de que os parâmetros do intercepto e das variáveis explicativas sejam estatisticamente iguais a zero?
- **f.** Reestime a regressão mantendo apenas as variáveis consideradas estatisticamente significativas. Interprete e compare os resultados com a equação anterior.

2. Com base no arquivo **acoes.dta**, que traz dados sobre os retornos dos papéis das empresas ACESITA e CESP listados na Bolsa de Valores de São Paulo, bem como o retorno do próprio índice Ibovespa ao longo de um período composto por 71 dias úteis, pede-se:

- **a.** Estime como a variação do retorno do Ibovespa impacta no retorno da empresa ACESITA.
- **b.** Interprete o nível de significância da reta de regressão e dos parâmetros individuais, bem como o coeficiente de determinação.
- **c.** Estime um novo modelo, desta vez com o retorno da empresa CESP como variável dependente. Interprete os resultados.
- **d.** Se o retorno do Ibovespa alcançar o patamar de 0,5%, quais serão os retornos previstos para as ações das empresas ACESITA e CESP? Além disso, quais são os intervalos de previsão para os retornos das ações com nível de confiança de 95%?

CAPÍTULO 5

Avaliação dos Modelos de Regressão

No Capítulo 4, estudamos a técnica de regressão linear utilizando os modelos simples e múltiplo. Neste capítulo, iremos aprofundar alguns conceitos relativos à avaliação dos modelos regressivos estimados, além de tratar da aplicação dos testes de hipóteses e da transformação de variáveis.

Usaremos em nossos exemplos a base de dados **paises.dta**. A referida base possui 79 observações sobre dados simulados relativos a países. É composta pelas variáveis contidas no Quadro 5.1.

Na janela de comandos do aplicativo Stata®, solicitaremos a abertura da base de dados **paises.dta**, utilizando o comando **use** (Resultados 5.1). Lembre-se de informar o endereço completo de localização do arquivo **paises.dta**.

RESULTADOS 5.1 Abertura do arquivo paises.dta.

```
. use "paises.dta"
(Dados simulados sobre países)
```

Quadro 5.1 Variáveis que compõem a base de dados **paises.dta**

Variável	Descrição	Tipo
pais	País	Qualitativa
pop	População	Quantitativa
nata	Taxa de natalidade	Quantitativa
mort	Taxa de mortalidade	Quantitativa
mor1	Mortalidade infantil (para crianças entre um a cinco anos)	Quantitativa
mor2	Mortalidade infantil (para crianças com até um ano)	Quantitativa
expe	Expectativa de vida	Quantitativa
pibp	PIB *per capita*	Quantitativa
urba	Percentual da população urbana	Quantitativa
esc1	Percentual da população com primeiro grau	Quantitativa
esc2	Percentual da população com segundo grau	Quantitativa

5.1. TESTES DE HIPÓTESES

Suponha que estamos interessados em conhecer a relação da taxa de natalidade (*nata*) nos países que compõem a amostra em função da expectativa de vida (*expe*) e percentual de pessoas com segundo grau (*esc2*).

Para realizar tal tarefa, iremos utilizar o seguinte comando:

regress nata expe esc2

Conforme vimos no Capítulo 4, o p-valor do teste F foi inferior a 0,0001, implicando a rejeição da hipótese nula de que todos os coeficientes estimados das variáveis explicativas são nulos. Individualmente, todos os p-valores dos testes t indicam que todas as variáveis explicativas e a constante foram consideradas significativas. O poder explicativo do modelo foi de aproximadamente 78,34% (Resultados 5.2).

RESULTADOS 5.2 Resultados da regressão múltipla.

```
. regress nata expe esc2

      Source |       SS       df       MS              Number of obs =      79
-------------+------------------------------           F(  2,    76) =  137.46
       Model |  5759.56004     2  2879.78002            Prob > F      =  0.0000
    Residual |  1592.17166    76  20.9496271            R-squared     =  0.7834
-------------+------------------------------           Adj R-squared =  0.7777
       Total |   7351.7317    78  94.2529705            Root MSE      =  4.5771

------------------------------------------------------------------------------
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.2912442    .109392    -2.66   0.009    -.5091173   -.0733711
        esc2 |  -.2487269   .0428599    -5.80   0.000    -.3340898    -.163364
       _cons |   58.06357    5.36357    10.83   0.000      47.3811    68.74605
------------------------------------------------------------------------------
```

O modelo estimado pode ser representado pela seguinte equação:

$$\text{estimação de } nata = 58{,}063 - 0{,}291.expe - 0{,}249.esc2 \qquad \text{[Equação 5.1]}$$

De acordo com o modelo anterior percebemos que, mantidas todas as demais variáveis inalteradas:

1. Países com maior expectativa de vida tendem a apresentar menor taxa de natalidade.
2. Países onde a maioria da população possui o segundo grau tendem a apresentar menor taxa de natalidade.

Agora, imaginemos que estamos interessados em comparar os parâmetros estimados com outros valores ou entre si. Para fazer a comparação de quaisquer valores em relação aos coeficientes estimados, o Stata® utiliza o teste de Wald, por meio do comando **test** (Sintaxe 5.1).

SINTAXE 5.1 Comando test.

test exp

Em que:

- exp: Expressão que será considerada como hipótese nula do teste.

No exemplo anterior, verificamos que o coeficiente estimado para a variável *expe* foi de -0,291. Supondo que, em uma pesquisa anterior, o coeficiente estimado tivesse sido de -0,34. Assim, iremos testar se o valor estimado pela regressão atual difere significativamente do obtido na regressão anterior. Devemos digitar o seguinte comando no Stata®:

test expe = -0.34

Verificamos que, mesmo que se considerasse um nível de significância de 10%, com um p-valor de 0,657, não haveria rejeição da hipótese nula do teste que, nesse caso, foi a seguinte: H_0: β_{expe} = -0,34 (Resultados 5.3).

RESULTADOS 5.3 Teste de Wald para os coeficientes de uma regressão.

```
. test expe = -0.34

 ( 1)  expe = -.34

       F(  1,    76) =    0.20
            Prob > F =    0.6571
```

Suponhamos que desejamos verificar se a intensidade do efeito da variável *expe* é a mesma da variável *esc2*. Novamente faremos uso do teste de Wald, informando na janela de comandos o seguinte:

test expe = esc2

Para verificar se os coeficientes das variáveis *expe* e *esc2* são iguais, o Stata® reconstruiu a expressão que informamos de modo a comparar se a mesma é igual a zero. Assim sendo, a hipótese nula que foi informada H_0: $\beta_{expe} = \beta_{esc2}$ foi modificada para H_0: $\beta_{expe} - \beta_{esc2} = 0$.

Com um p-valor superior a 0,7, concluímos que não houve rejeição da hipótese nula, e que, em módulo, as variáveis *expe* e *esc2* afetam a taxa de natalidade com a mesma intensidade, do ponto de vista estatístico (Resultados 5.4).

RESULTADOS 5.4 Teste de Wald para os coeficientes de uma regressão.

```
. test expe = esc2

 ( 1)  expe - esc2 = 0

       F(  1,    76) =    0.08
            Prob > F =    0.7749
```

Por último, imaginemos que, em outro estudo, foi identificado que a soma dos coeficientes das variáveis *esc2* e *expe* foi igual -0,9. Para testar se a situação se repetiu na presente regressão, utilizaremos o seguinte comando:

test esc2 + expe == -0.9

Com um p-valor inferior a 0,0001 no teste de Wald, considerando qualquer um dos níveis de significância usuais, rejeitamos a hipótese nula de que, na nova regressão, a soma desses coeficientes seja igual a -0,9 (Resultados 5.5).

RESULTADOS 5.5 Teste de Wald para os coeficientes de uma regressão.

```
. test esc2 + expe == -0.9

 ( 1)  expe + esc2 = -.9

       F(  1,    76) =   22.90
            Prob > F =    0.0000
```

Para acessar o teste de Wald, após uma regressão, via barra de menus, precisamos selecionar as seguintes opções: *Statistics* → *Postestimation* → *Tests* → *Test linear hypotheses*. Surgirá uma janela, conforme a Figura 5.1.

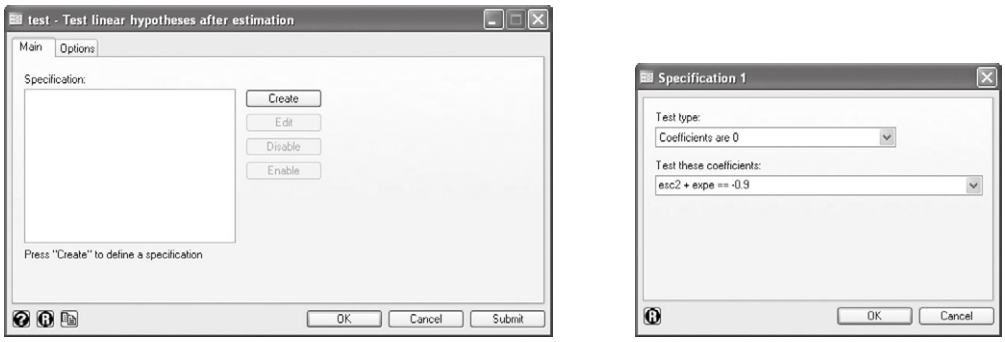


Figura 5.1 ***Janelas de configurações do comando*** **test.**

5.2. MULTICOLINEARIDADE

A multicolinearidade ocorre quando duas ou mais variáveis explicativas possuem correlação entre si. Quando a multicolinearidade se dá em um grau bastante elevado, podem ser gerados vieses bastante expressivos nos parâmetros estimados em uma regressão.

Conforme vimos no Capítulo 4, não há um teste amplamente aceito para a detecção da multicolinearidade. Para detectar a sua presença, costumamos utilizar algumas estatísticas, tais como a correlação linear e o fator de inflação da variância ou VIF (*variance inflation factor*).

No exemplo a ser utilizado, queremos analisar a relação da taxa de mortalidade com as seguintes variáveis explicativas: *mor1*, *mor2* e *expe*.

Inicialmente, iremos solicitar a correlação linear entre essas variáveis, utilizando o seguinte comando:

pwcorr mort mor1 mor2 expe, sig

Observando os Resultados 5.6, percebemos que todas as variáveis explicativas possuem correlações, entre si, superiores a 0,8 a um nível de significância de 1%. Variáveis

RESULTADOS 5.6 Análise da correlação entre variáveis.

```
. pwcorr mort mor1 mor2 expe, sig

             |     mort     mor1     mor2     expe
-------------+------------------------------------
        mort |   1.0000
             |
             |
        mor1 |   0.4395   1.0000
             |   0.0001
             |
        mor2 |   0.4735   0.9895   1.0000
             |   0.0000   0.0000
             |
        expe |  -0.5610  -0.9096  -0.8938   1.0000
             |   0.0000   0.0000   0.0000
```

explicativas fortemente correlacionadas são um forte indicativo de que haverá problemas de multicolinearidade.

Agora, solicitaremos a estimação dos parâmetros da regressão. Utilizaremos o comando **regress** em sua forma reduzida (**reg**), informando o seguinte:

reg mort mor1 mor2 expe

Verificamos que os resultados, tanto do teste F quanto do teste t, indicam que as variáveis explicativas possuem coeficientes estatisticamente significativos. Os R^2 e R^2 Ajustado alcançaram os valores de 0,439 e 0,416, respectivamente (Resultados 5.7). Passaremos para a análise das estatísticas VIF.

RESULTADOS 5.7 Resultados da regressão múltipla.

```
. reg mort mor1 mor2 expe

      Source |       SS       df       MS              Number of obs =      79
-------------+------------------------------           F(  3,    75) =   19.56
       Model |  425.989467     3  141.996489            Prob > F      =  0.0000
    Residual |  544.603811    75  7.26138414            R-squared     =  0.4389
-------------+------------------------------           Adj R-squared =  0.4165
       Total |  970.593278    78  12.4435036            Root MSE      =  2.6947

------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mor1 |  -.3475482   .0866649    -4.01   0.000    -.5201936   -.1749028
        mor2 |   .1770391   .0496353     3.57   0.001     .0781605    .2759177
        expe |  -.3733709   .0772901    -4.83   0.000    -.5273406   -.2194012
       _cons |   37.66457   6.132615     6.14   0.000     25.44777    49.88136
------------------------------------------------------------------------------
```

Devemos digitar o seguinte na janela de comandos do Stata®:

estat vif

De acordo com as estatísticas VIF, observamos que, com exceção da variável relativa à expectativa de vida, as demais variáveis explicativas apresentaram estatísticas superiores a 10 (Resultados 5.8). Segundo Gujarati (2011), um VIF acima de 10 é indicativo de multicolinearidade, porém, Fávero *et al.* (2009) argumentam que um VIF acima de 5 já pode causar problemas de multicolinearidade, conforme já discutido no Capítulo 4.

Em razão de tais resultados, podemos concluir que o modelo estimado apresenta problemas de multicolinearidade que podem enviesar os parâmetros estimados.

RESULTADOS 5.8 Estatísticas VIF.

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
        mor1 |     56.22    0.017788
        mor2 |     48.25    0.020725
        expe |      5.86    0.170727
-------------+----------------------
    Mean VIF |     36.78
```

5.3. HETEROCEDASTICIDADE

No Capítulo 4 foram apresentados os pressupostos do estimador de mínimos quadrados utilizados pelo Stata® no comando **regress**, para as regressões lineares simples e múltiplas. Dentre os pressupostos, está definido que os resíduos devem ser homocedásticos, ou seja, não devem haver problemas de heterocedasticidade.

O teste para a detecção da heterocedasticidade foi o Breusch-Pagan, executado no Stata® por intermédio do comando **estat hettest** ou simples **hettest** (principalmente nas versões mais antigas). Apresentamos novamente a sintaxe deste comando, incluindo novas opções (Sintaxe 5.2).

SINTAXE 5.2 Comando estat hettest.

estat hettest [varlist] [, iid] [, fstat]

Em que:

- varlist: Lista contendo as variáveis explicativas que serão utilizadas no cômputo da estatística do teste. Caso não seja informada nenhuma variável, o Stata® utilizará as variáveis explicativas da última regressão estimada.
- iid: Utiliza a estatística NR2, no lugar da estatística-padrão do teste.
- fstat: Utiliza a estatística F, no lugar da estatística-padrão do teste.

Voltaremos a realizar a estimativa do primeiro modelo, que tem como variável dependente a taxa de natalidade. Depois solicitaremos o teste Breusch-Pagan. Devemos informar os seguintes comandos:

reg nata expe esc2
hettest

Com um p-valor de 0,0101, concluímos, de acordo com o teste Breusch-Pagan, que a hipótese nula foi rejeitada (Resultados 5.9). Dessa forma, os resíduos da regressão são considerados heterocedásticos.

RESULTADOS 5.9 Resultados da regressão múltipla e teste Breusch-Pagan.

```
. reg nata expe esc2

      Source |       SS       df       MS              Number of obs =      79
-------------+------------------------------           F(  2,    76) =  137.46
       Model |  5759.56004     2  2879.78002            Prob > F      =  0.0000
    Residual |  1592.17166    76  20.9496271            R-squared     =  0.7834
-------------+------------------------------           Adj R-squared =  0.7777
       Total |   7351.7317    78  94.2529705            Root MSE      =  4.5771

------------------------------------------------------------------------------
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.2912442    .109392    -2.66   0.009    -.5091173   -.0733711
        esc2 |  -.2487269   .0428599    -5.80   0.000    -.3340898    -.163364
       _cons |   58.06357    5.36357    10.83   0.000     47.3811    68.74605
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of nata

         chi2(1)      =     6.62
         Prob > chi2  =   0.0101
```

As opções do comando **estat hettest** somente deverão ser utilizadas quando o estimador utilizado na regressão não tiver como pressuposto que os resíduos possuem distribuição normal. Não é o caso do estimador dos mínimos quadrados.

Existe no Stata® outro teste para detecção de heterocedasticidade: o teste de White. Esse teste é executado por meio do comando **estat imtest** ou simplesmente **imtest** (especialmente nas versões mais antigas) (Sintaxe 5.3).

SINTAXE 5.3 Comando estat imtest.

estat imtest, white

Em que:

- white: Essa opção deve ser informada para que o Stata® realize o teste de White no formato original.

Continuando com o nosso exemplo, informaremos, na janela de comandos, o seguinte:

imtest, white

O teste de White possui hipóteses semelhantes às do teste Breusch-Pagan, isto é: H_0: os resíduos são homocedásticos, e H_1: os resíduos são heterocedásticos. Verificamos que o teste de White também indicou que os resíduos são heterocedásticos em razão da rejeição da hipótese nula (Resultados 5.10).

RESULTADOS 5.10 Teste de White.

```
. imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     15.93
         Prob > chi2  =    0.0070

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |      15.93      5    0.0070
            Skewness |       5.88      2    0.0528
            Kurtosis |       0.49      1    0.4856
---------------------+-----------------------------
               Total |      22.30      8    0.0044
---------------------------------------------------
```

Para acessar o teste Breusch-Pagan, utilizando a barra de menus, devemos clicar nas seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Aparecerá uma janela, conforme a Figura 5.2.

Para acessar o teste de White, utilizando a barra de menus, devemos clicar nas seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Irá aparecer uma janela, conforme a Figura 5.3.

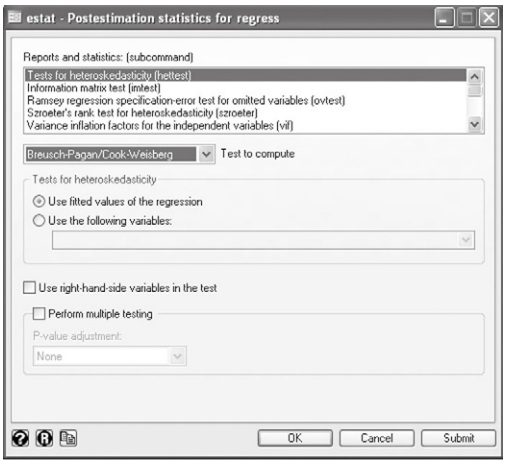


Figura 5.2 *Janela de configurações do comando* **estat** *selecionando-se a opção* **hettest.**

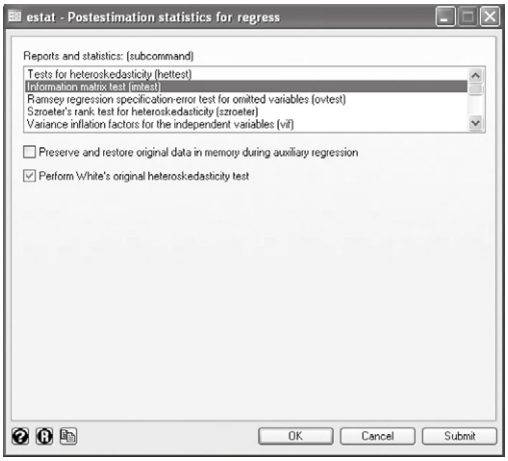


Figura 5.3 *Janela de configurações do comando* **estat** *selecionando-se a opção* **imtest.**

5.4. GRÁFICOS DE AVALIAÇÃO

Podemos utilizar alguns dos recursos gráficos para avaliar as estimações realizadas em uma regressão.

Alguns gráficos podem ser empregados para a detecção de *outliers*, utilizando-se séries obtidas a partir do comando **predict**, após uma regressão. Antes de analisarmos esses gráficos, vamos estudar a sintaxe de novas opções para o comando **predict** (Sintaxe 5.4).

O primeiro gráfico que iremos obter é o histograma. Para gerar este gráfico utilizaremos o comando **histogram**, conforme vimos no Capítulo 2.

SINTAXE 5.4 Comando predict.

predict newvar [, rstudent]

Em que:

- newvar: Nome da nova variável que armazenará os valores previstos.
- rstudent: Opção a ser utilizada para a geração dos resíduos estudentizados da regressão.

Para a identificação de *outliers*, iremos utilizar os resíduos estudentizados da regressão e exibi-los no histograma da série. Na janela de comandos do Stata®, digitaremos os seguintes comandos:

```
predict res1, rstudent
histogram res1
```

RESULTADOS 5.11 Gerando o histograma dos resíduos estudentizados.

```
. predict res1, rstudent

. histogram res1
(bin=8, start=-2.2365158, width=.61626831)
```

Após a análise do histograma dos resíduos estudentizados (Figura 5.4), verificamos que existem observações cujos resíduos foram superiores a dois em módulo, sendo provável

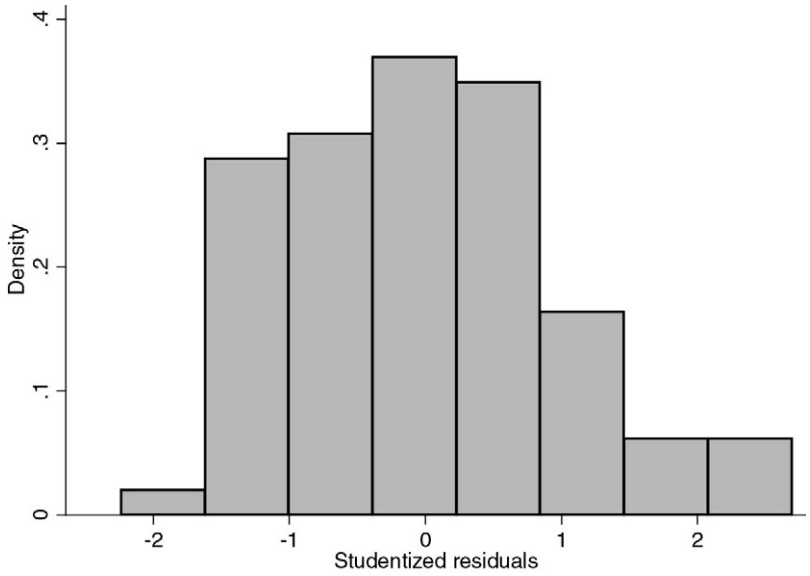


Figura 5.4 ***Histograma dos resíduos estudentizados.***

a existência de *outliers*. Para localizarmos essas observações, iremos utilizar o comando **list**, informando na janela de comandos o seguinte:

list pais nata expe esc2 res1 if abs(res1) > 2

Os países de códigos 33, 43, 46 e 73 apresentaram resíduos com valores acima de dois em módulo (Resultados 5.12). Em uma rápida análise, podemos perceber que a taxa de natalidade do país de código 43 é relativamente mais baixa do que a dos outros países, quando comparada com expectativa de vida similar (país de código 46).

RESULTADOS 5.12 Listando possíveis *outliers* em função dos resíduos estudentizados.

```
. list pais nata expe esc2 res1 if abs(res1) > 2

     +-------------------------------------------------+
     | pais     nata        expe       esc2       res1 |
     |-------------------------------------------------|
 33. |   33     21.5   81.553659   98.46043   2.693631 |
 43. |   43   28.069   46.669366   28.03876  -2.236516 |
 46. |   46   46.914   50.536049   27.63603    2.44587 |
 73. |   73   27.923   66.967683   84.04336   2.378645 |
     +-------------------------------------------------+
```

Para acessar o comando **predict**, por meio da barra de menus, basta selecionar as seguintes opções: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Será exibida uma janela, conforme a Figura 5.5.

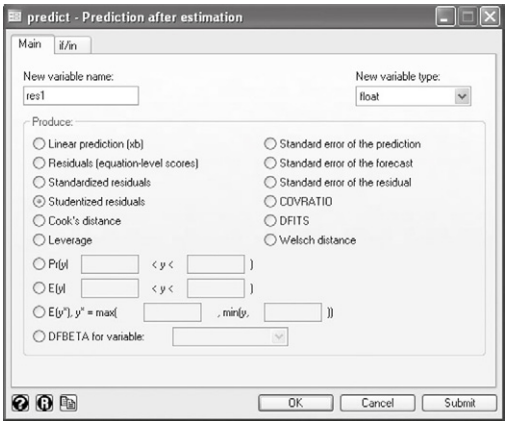


Figura 5.5 *Janela de configurações do comando* **predict.**

Por intermédio da barra de menus, acessamos o comando **histogram**, por meio das seguintes opções: *Graphics* → *Histogram*. Será exibida uma janela, conforme a Figura 5.6.

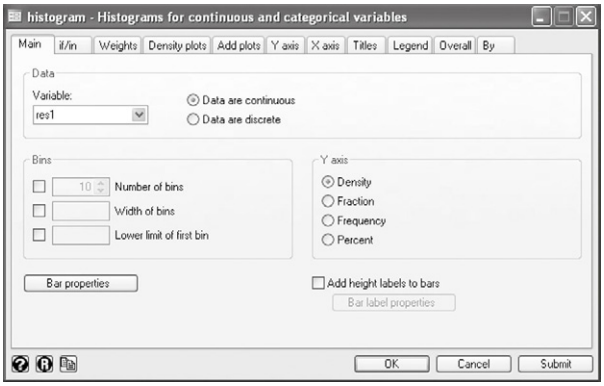


Figura 5.6 *Janela de configurações do comando* **histogram.**

Outra forma de identificação de *outliers* é o uso do gráfico de dispersão das distâncias de *leverage*. No Stata®, a geração deste gráfico é possível por meio do comando **lvr2plot** (Sintaxe 5.5).

SINTAXE 5.5 Comando lvr2plot.

lvr2plot [, mlabel(varname)]

Em que:

- varname: Nome da variável que será utilizada para rotular os pontos no gráfico.

Vamos agora verificar o gráfico de dispersão das distâncias de *leverage*. Inicialmente iremos criar um índice para que possamos identificar os pontos no gráfico e, na sequência, iremos solicitar a geração do gráfico (Figura 5.7). Para tanto, precisamos digitar o seguinte comando no Stata®:

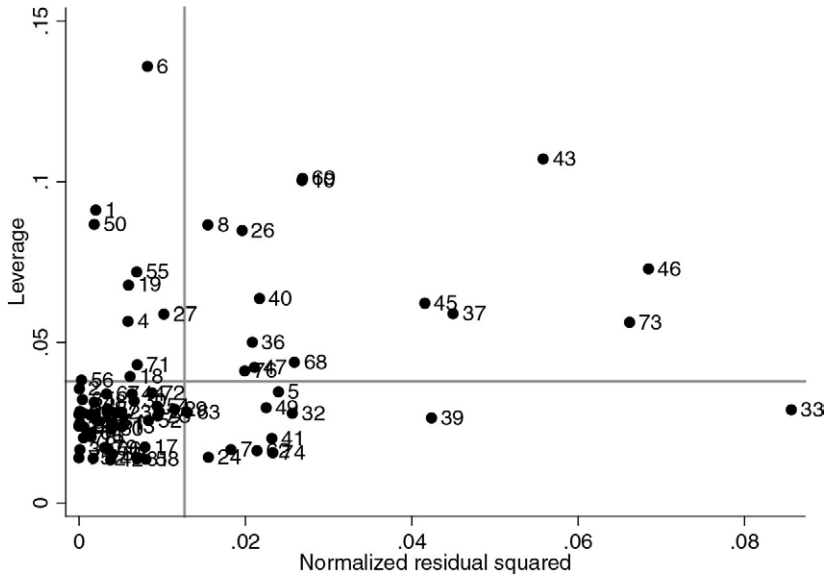


Figura 5.7 ***Gráfico de dispersão das distâncias de leverage.***

lvr2plot, mlabel(pais)

RESULTADOS 5.13 Gerando o gráfico de dispersão das distâncias de *leverage*.

```
lvr2plot, mlabel(pais)
```

Na análise gráfica, podemos notar que as observações 6 e 43 são aquelas que apresentam as maiores distâncias de *leverage*, ou seja, podem ser consideradas como prováveis *outliers*. Para verificar as observações destacadas, usaremos, novamente, o comando **list**. Digitaremos o seguinte:

list pais nata expe esc2 res1 if pais == 6 | pais == 43

Apenas o país de código 43 havia sido identificado, quando observarmos o critério anterior. O país de código 6 não fora identificado, pois o respectivo resíduo foi inferior a dois (Resultados 5.14).

RESULTADOS 5.14 Listando possíveis *outliers* em função das distâncias de *leverage*.

```
. list pais nata expe esc2 res1 if pais == 6 | pais == 43

     +-------------------------------------------------+
     | pais     nata        expe        esc2        res1 |
     |-------------------------------------------------|
  6. |    6   23.814   53.011537    61.07163   -.8492496 |
 43. |   43   28.069   46.669366    28.03876   -2.236516 |
     +-------------------------------------------------+
```

Para acessar o comando **lvr2plot**, é necessário selecionar as seguintes opções: *Statistics* → *Linear models and related* → *Regression diagnostics* → *Leverage-versus-squared-residual plot*. Aparecerá uma janela, conforme a Figura 5.8.

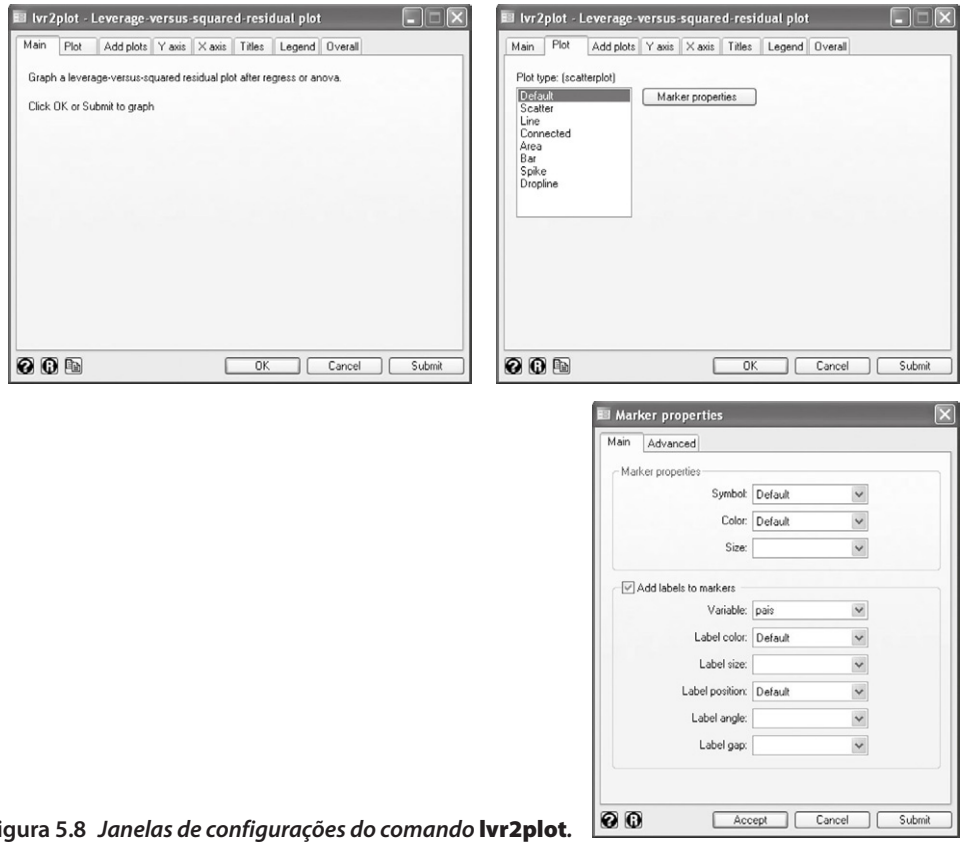


Figura 5.8 *Janelas de configurações do comando* **lvr2plot.**

Para acessar o comando **list**, é necessário selecionar as seguintes opções: *Data* → *Describe data* → *List data*. Irá aparecer uma janela, conforme a Figura 5.9.

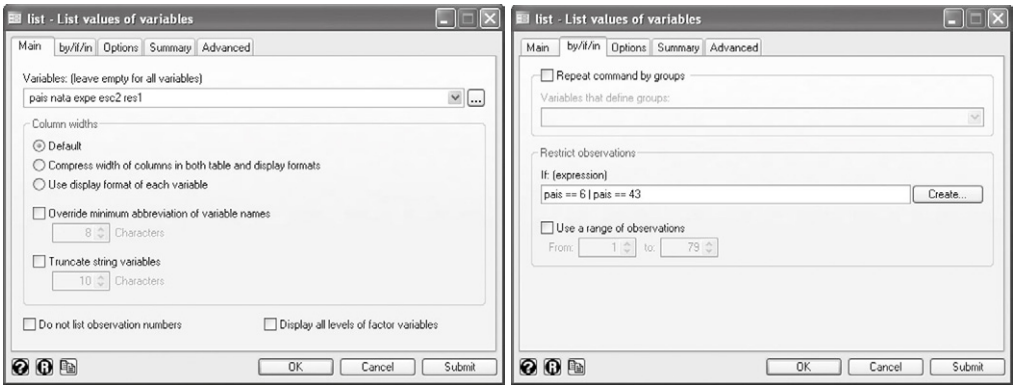


Figura 5.9 ***Janelas de configurações do comando*** **list.**

5.5. TRANSFORMAÇÃO DE VARIÁVEIS

Às vezes são necessárias algumas transformações em variáveis para evitar ou amenizar problemas ocasionados em uma regressão. Diferenças de escala, excesso de assimetria e excesso de curtose são apenas alguns exemplos de características de uma variável que podem torná-la problemática em uma estimação.

Primeiramente, iremos observar o histograma da variável *pop*. Informaremos, na janela de comandos, o seguinte:

histogram pop

RESULTADOS 5.15 Gerando o histograma da variável *pop*

```
. histogram pop
(bin=10, start=1, width=103.93)
```

No histograma (Figura 5.10) podemos observar que a variável *pop* é assimétrica positiva, isto é, há grande concentração de valores abaixo da média da variável.

A primeira transformação que faremos se refere à logaritmização de uma variável. O logaritmo de uma variável permite que seja preservada a relação linear entre duas variáveis e reduz problemas de assimetria.

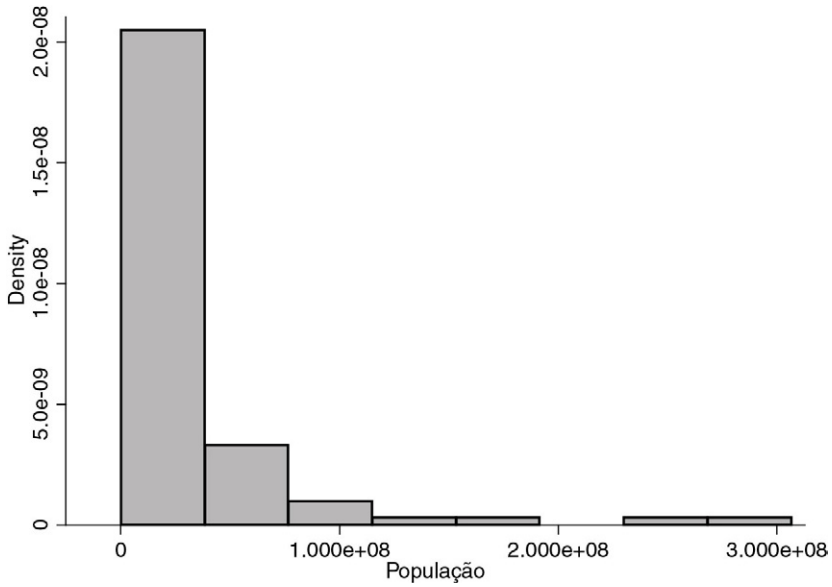


Figura 5.10 ***Histograma da variável*** **pop.**

Utilizaremos o comando **gen** para criar o logaritmo da variável *pop*. O Stata® emprega a função **log** para criar o logaritmo natural de uma variável. Informaremos o seguinte na janela de comandos:

gen lpop = log(pop)
histogram lpop

RESULTADOS 5.16 Criando o logaritmo da variável *pop* e gerando o histograma da variável *lpop*.

```
. gen lpop = log(pop)

. histogram lpop
(bin=8, start=11.601568, width=.99250579)
```

Visualmente, verificamos que o histograma da nova variável é menos assimétrico do que o histograma da variável original (Figura 5.11).

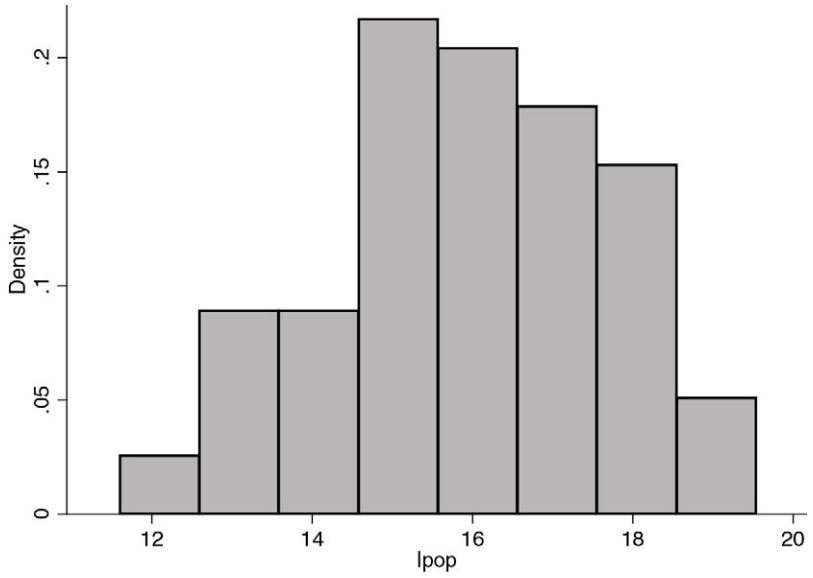


Figura 5.11 ***Histograma da variável*** **lpop.**

Outra transformação possível é a transformação de Box-Cox, que busca resolver problemas de assimetria, tornando a distribuição da variável a mais simétrica possível. No Stata®, tal transformação é elaborada por meio do comando **bcskew0** (Sintaxe 5.6).

SINTAXE 5.6 Comando bcskew0.

bcskew0 newvar = varname

Em que:

- newvar: Nome da variável que será criada.
- varname: Nome da variável que será transformada.

Agora, digitaremos os seguintes comandos:

bcskew0 bpop = pop

histogram bpop

De acordo com o resultado da transformação de Box-Cox e com o gráfico da nova variável, verificamos que se trata de uma distribuição cuja medida de assimetria é de 0,0001, o que nos leva a considerar tal distribuição como simétrica (Resultados 5.17 e Figura 5.12).

RESULTADOS 5.17 Utilizando a transformação de Box-Cox na variável *pop* e gerando o histograma da variável *bpop*

```
. bcskew0 bpop = pop

       Transform |          L     [95% Conf. Interval]        Skewness
-----------------+---------------------------------------------------
     (popˆL-1)/L |   .0551868       (not calculated)          .0000174

. histogram bpop
(bin=8, start=16.253193, width=2.3627157)
```

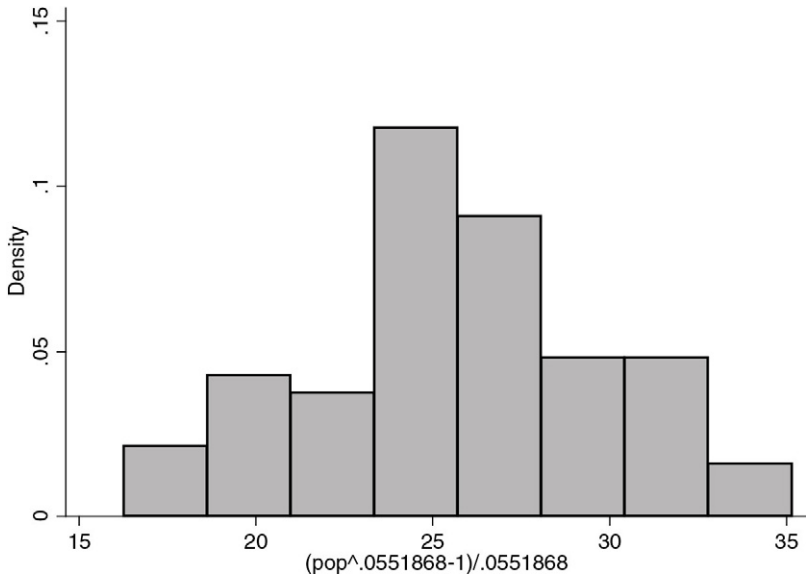


Figura 5.12 ***Histograma da variável*** **bpop.**

Para observar o impacto dessas transformações nas relações entre as variáveis *nata* e *pop*, iremos solicitar as correlações por intermédio do seguinte comando:

pwcorr nata pop lpop bpop, sig

A variável *pop* não apresenta correlação significativa com a variável *nata*. Possivelmente, a assimetria excessiva da variável original é a principal responsável por tal situação. Quando comparamos as variáveis transformadas, vemos que ambas, apesar de não apresentarem correlações significativas, possuem maior correlação com a variável *nata* do que com a variável original (Resultados 5.18).

RESULTADOS 5.18 Correlações entre as variáveis.

```
. pwcorr nata pop lpop bpop, sig

             |     nata      pop     lpop     bpop
-------------+------------------------------------
        nata |   1.0000
             |
             |
         pop |  -0.0529   1.0000
             |   0.6431
             |
        lpop |   0.0852   0.6978   1.0000
             |   0.4552   0.0000
             |
        bpop |   0.0756   0.7329   0.9982   1.0000
             |   0.5081   0.0000   0.0000
             |
```

Para acessar a transformação de Box-Cox, via barra de menus, devemos clicar nas seguintes opções: *Data* → *Create or change data* → *Other variable-creation commands* → *Box -Cox transform*. Será exibida uma janela, conforme a Figura 5.13.

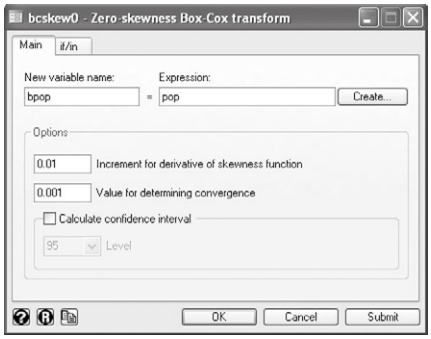


Figura 5.13 ***Janela de configurações do comando*** **bcskew0.**

O comando **bcskew0** (transformação de Box-Cox com imposição de assimetria nula para uma nova variável *bvar*) faz com que seja gerado um parâmetro *L* tal que esta nova variável se relacione com a variável original (*var*) por meio da seguinte expressão:

$$b\,\mathrm{var} = \frac{\left(\mathrm{var}^{L} - 1\right)}{L} \qquad \text{[Equação 5.2]}$$

O comando **bcskew0** é muito utilizado para os casos em que a variável dependente de um modelo de regressão não apresenta distribuição normal, o que fere o primeiro

pressuposto da estimação pelo método dos mínimos quadrados ordinários. Neste caso, uma nova variável pode ser gerada a partir da variável original, a fim de que eventualmente possa ser verificado o pressuposto da normalidade da variável dependente do modelo, mesmo que este passe a ter uma diferente forma funcional. Cabe ao pesquisador definir a melhor forma funcional do modelo a ser utilizado, em função da teoria subjacente e da sua experiência, respeitando-se os pressupostos da estimação.

5.6. EXERCÍCIOS

1. O arquivo **salarios.dta** apresenta dados sobre os salários de 15 alunos recém-formados no curso de Administração de empresas. Traz também três exemplos que contêm, cada um deles, as notas finais de RH e de econometria (de 0 a 10) que estes alunos tiraram na faculdade. Pede-se:
   a. Para cada um dos exemplos propostos, elabore o modelo de regressão linear múltipla *salário = f (nota de RH; nota de econometria)*.
   b. Após elaborar cada um dos três modelos, interprete os *outputs* com foco para o teste F e os testes t. Há alguma inconsistência quando da análise destes *outputs*?
   c. Elabore a matriz de correlações para as variáveis *RH* e *econometria* em cada um dos casos. As correlações são muito altas, porém, diferentes de 1, em algum dos três casos? Se sim, como você interpretaria este fenômeno?
   d. Elabore e discuta as estatísticas VIF para cada um dos três modelos.
2. Por meio do arquivo **Renda x Tempo Formado.dta**, elabore o modelo de regressão linear simples *renda = f (tempo de formado)* e discuta a existência de heterocedasticidade no modelo. Elabore um gráfico de dispersão de *renda = f (tempo formado)* para auxiliar na discussão.

CAPÍTULO 6

Regressão Robusta

A regressão robusta é um método alternativo ao método dos mínimos quadrados quando existem *outliers* e opta-se pela sua manutenção na análise. Além disso, também pode ser utilizado para detectar pontos de influência. O objetivo do presente capítulo é mostrar como aplicar vários comandos para a análise de dados com a presença de *outliers* em modelos de regressão.

Continuaremos a utilizar, em nosso exemplo, a base de dados **paises.dta**. A referida base possui 79 observações sobre dados simulados relativos a países. É composta pelas variáveis descritas no Quadro 6.1.

Na janela de comandos do aplicativo Stata® solicitaremos a abertura da base de dados **paises.dta**, utilizando o comando **use** (Resultados 6.1). Lembre-se de informar o endereço completo de localização do arquivo **paises.dta**.

RESULTADOS 6.1 Abertura do arquivo paises.dta.

```
. use "paises.dta"
(Dados simulados sobre países)
```

Quadro 6.1 Variáveis que compõem a base de dados **paises.dta**

Variável	Descrição	Tipo
pais	País	
pop	População	Quantitativa
nata	Taxa de natalidade	Quantitativa
mort	Taxa de mortalidade	Quantitativa
mor1	Mortalidade infantil (para criança entre um e cinco anos)	Quantitativa
mor2	Mortalidade infantil (para criança com até um ano)	Quantitativa
expe	Expectativa de vida	Quantitativa
pibp	PIB *per capita*	Quantitativa
urba	Percentual da população urbana	Quantitativa
esc1	Percentual da população com primeiro grau	Quantitativa
esc2	Percentual da população com segundo grau	Quantitativa

6.1. OUTLIERS

Na regressão linear, os resíduos consistem na diferença entre o valor previsto (baseado na equação da regressão) e o valor observado. Na regressão linear, um *outlier* pode indicar uma observação com altos valores dos resíduos, em decorrência de uma peculiaridade da amostra ou um erro na digitação dos dados.

No Capítulo 5 começamos a verificar algumas análises gráficas para a detecção de *outliers*. Agora, procedemos no sentido de ampliar a lista de procedimentos utilizados para tal tarefa.

Suponha que o nosso objetivo seja entender quais condições seriam capazes de explicar a taxa de mortalidade infantil (para crianças com menos de um ano de idade), utilizando as características dos países.

Inicialmente, estimaremos uma regressão linear múltipla (Resultados 6.2), com o comando **reg**. Digitaremos o seguinte na janela de comandos:

reg nata expe esc2

RESULTADOS 6.2 Resultados da regressão múltipla.

```
. reg nata expe esc2

      Source |       SS       df       MS              Number of obs =      79
-------------+------------------------------           F(  2,    76) =  137.46
       Model |  5759.56004     2  2879.78002            Prob > F      =  0.0000
    Residual |  1592.17166    76  20.9496271            R-squared     =  0.7834
-------------+------------------------------           Adj R-squared =  0.7777
       Total |   7351.7317    78  94.2529705            Root MSE      =  4.5771

------------------------------------------------------------------------------
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.2912442    .109392    -2.66   0.009    -.5091173   -.0733711
        esc2 |  -.2487269   .0428599    -5.80   0.000    -.3340898    -.163364
       _cons |   58.06357    5.36357    10.83   0.000      47.3811    68.74605
------------------------------------------------------------------------------
```

Todos os procedimentos para a detecção de *outliers* dependerá de estatísticas que serão preditas após a estimação de uma regressão. O comando utilizado é o **predict**, já estudado, porém agora o apresentaremos com novas opções (Sintaxe 6.1).

A primeira medida que utilizaremos é a distância de *leverage*, que mensura o quanto uma observação influencia os coeficientes de uma regressão. Uma observação pode ser considerada como *outlier* se a distância de *leverage* for maior que $2 \cdot k / N$, em que k é o número de parâmetros (incluindo o intercepto) e N é o tamanho da amostra. Pontos com distâncias elevadas podem apresentar um grande efeito na estimação dos coeficientes da regressão.

SINTAXE 6.1 Comando predict.

predict newvar [, leverage] [, cooksd] [, difts] [, covratio]

Em que:

- newvar: Nome da nova variável que armazenará os valores previstos.
- leverage: Opção a ser utilizada para a geração das distâncias de *leverage*.
- cooksd: Opção a ser utilizada para a geração das distâncias de Cook.
- difts: Opção a ser utilizada para a geração do indicador DfFit.
- covratio: Opção a ser utilizada para a geração do indicador de covariância.

A distância de *leverage* varia de 0 a 1.Valores próximos de 1 ou superiores a 0,5 podem indicar problemas. No Stata®, digitaremos o seguinte comando:

predict lev, leverage

RESULTADOS 6.3 Gerando as distâncias de *leverage*.

```
. predict lev, leverage
```

Agora que já temos as distâncias, precisamos calcular o valor crítico que nos orientará na detecção dos *outliers*. Para tanto, utilizaremos o comando **display**, que possui a seguinte sintaxe (Sintaxe 6.2).

SINTAXE 6.2 Comando display.

display exp

Em que:

- exp: Expressão que será calculada ou exibida na janela de resultados.

Informaremos no Stata® o seguinte:

display 2 * 4 / 79

RESULTADOS 6.4 Exibindo o valor crítico para comparar as distâncias de *leverage*.

```
. display 2 * 4 / 79
.10126582
```

Verificamos que o valor crítico a ser utilizado é 0,101, com aproximação. As observações com distâncias de *leverage*, acima do valor crítico, serão consideradas como *outliers*. Para identificar se há observações nessa situação, iremos utilizar o comando **list** da seguinte forma:

list pais mor2 nata esc1 esc2 lev if lev > 0.101

Empregando esse critério verificamos a existência de duas observações, que podem ser consideradas como *outliers*: 6 e 43 (Resultados 6.5).

RESULTADOS 6.5 Detectando *outliers* utilizando as distâncias de *leverage*.

```
. list pais mor2 nata esc1 esc2 lev if lev > 0.101

     +-------------------------------------------------------+
     | pais   mor2     nata       esc1       esc2        lev |
     |-------------------------------------------------------|
  6. |    6   30.1   23.814   87.07093   61.07163   .1358258 |
 43. |   43   95.8   28.069   71.92008   28.03876   .1070178 |
     +-------------------------------------------------------+
```

Para acessar o comando **predict**, precisamos selecionar os seguintes comandos na barra de menus: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Aparecerá a tela da Figura 6.1.

O comando **display** pode ser acessado, via barra de menus, clicando-se nas seguintes opções: *Data* → *Other utilities* → *Hand calculator* (Figura 6.2).

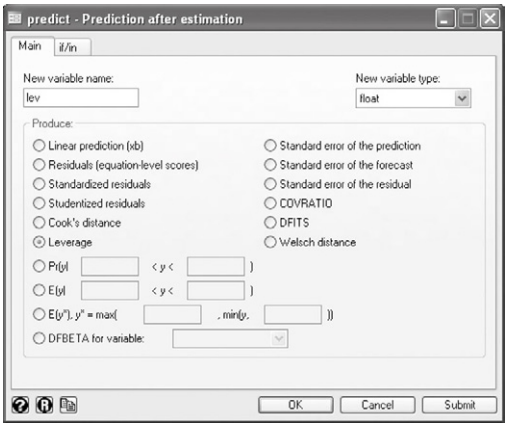


Figura 6.1 *Janela de configurações do comando* **predict** *selecionando-se a opção* **Leverage.**

Figura 6.2 *Janela de configurações do comando* **display.**

A distância de Cook, outra medida utilizada para a detecção de *outliers*, combina informações da distância de *leverage* e dos resíduos da observação. Mede o quanto uma observação influencia o modelo global ou os valores previstos.

Uma observação é considerada de grande influência se a distância de Cook é maior do que 4 / N, em que N é o tamanho da amostra. Assim, uma distância maior do que 1 indica um grande problema de *outlier*.

No Stata®, iremos utilizar os seguintes comandos:

predict cook, cooksd

display 4 / 79

RESULTADOS 6.6 Gerando as distâncias de Cook e calculando o valor crítico.

```
. predict cook, cooksd

. display 4 / 79
.05063291
```

Para verificar a existência de observações cuja distância de Cook seja superior a 0,051, iremos utilizar o seguinte comando:

list pais mor2 nata esc1 esc2 cook if cook > 0.051

Caso optássemos pela distância de Cook para o procedimento de detecção de *outliers*, identificaríamos um total de oito observações: 10, 33, 37, 43, 45, 46, 69 e 73 (Resultados 6.7).

Para gerar as distâncias de Cook, precisamos selecionar os seguintes comandos na barra de menus: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Surgirá uma tela, conforme a Figura 6.3.

RESULTADOS 6.7 Detectando *outliers* utilizando as distâncias de Cook.

```
. list pais mor2 nata esc1 esc2 cook if cook > 0.051

     +-------------------------------------------------------------+
     | pais    mor2     nata      esc1      esc2       cook |
     |-------------------------------------------------------------|
 10. |   10   165.7   35.221   68.05666    10.655   .0842769 |
 33. |   33     4.7     21.5   96.90646   98.46043    .066562 |
 37. |   37    79.4   37.824    82.7824   50.02514   .0759514 |
 43. |   43    95.8   28.069   71.92008   28.03876   .1898229 |
 45. |   45    95.4    44.16   96.89612   26.92535   .0743705 |
     |-------------------------------------------------------------|
 46. |   46   182.1   46.914   61.13494   27.63603   .1470478 |
 69. |   69   115.3   29.652   81.84889    31.8434   .0851279 |
 73. |   73    68.2   27.923   96.88607   84.04336   .1057143 |
     +-------------------------------------------------------------+
```

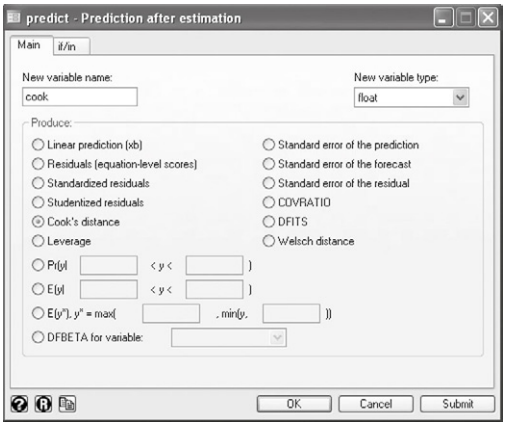


Figura 6.3 ***Janela de configurações do comando*** **predict** ***selecionando-se a opção*** **Cook's distance.**

O DfFit é o indicador de alavancagem e de resíduos elevados. É outra medida que pode ser utilizada para a detecção de *outliers*. Mensura o quanto uma observação influencia o modelo de regressão como um todo e o quanto os valores previstos são alterados pela inclusão ou exclusão de uma observação particular.

Uma observação é considerada *outlier* se $|\text{DfFIT}| > 2 \cdot \text{SQRT}(k/N)$, em que k é o número de parâmetros (incluindo o intercepto) e N é o tamanho da amostra (SQRT = raiz quadrada).

Na janela de comandos do Stata®, digitaremos o seguinte:

predict dfits, dfits
display 2 * sqrt(4 / 79)

RESULTADOS 6.8 Gerando o indicador DfFit e calculando o valor crítico.

```
. predict dfits, dfits

. display 2 * sqrt(4 / 79)
.45003516
```

Para verificar a existência de observações cujo indicador DfFit, em módulo, seja superior a 0,450, iremos utilizar o seguinte comando:

list pais mor2 nata esc1 esc2 dfits if abs(dfits) >0.450

De acordo com esse critério, oito observações foram consideradas como *outliers*. Os mesmos países então identificados quando empregamos as distâncias de Cook, também o foram com o indicador DfFit (Resultados 6.9).

RESULTADOS 6.9 Detectando *outliers* utilizando o indicador DfFit.

```
. list pais mor2 nata esc1 esc2 dfits if abs(dfits) > 0.450

     +-------------------------------------------------------------+
     | pais    mor2     nata       esc1       esc2       dfits |
     |-------------------------------------------------------------|
 10. |   10   165.7   35.221   68.05666     10.655   -.5071271 |
 33. |   33     4.7     21.5   96.90646   98.46043    .4648895 |
 37. |   37    79.4   37.824    82.7824   50.02514    .4859554 |
 43. |   43    95.8   28.069   71.92008   28.03876   -.7742453 |
 45. |   45    95.4    44.16   96.89612   26.92535    .4799896 |
     |-------------------------------------------------------------|
 46. |   46   182.1   46.914   61.13494   27.63603    .6856114 |
 69. |   69   115.3   29.652   81.84889    31.8434   -.5097016 |
 73. |   73    68.2   27.923   96.88607   84.04336    .5801554 |
     +-------------------------------------------------------------+
```

Para gerar o indicador DfFit, via barra de menus, devemos selecionar as seguintes opções: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Será exibida uma tela, conforme a Figura 6.4.

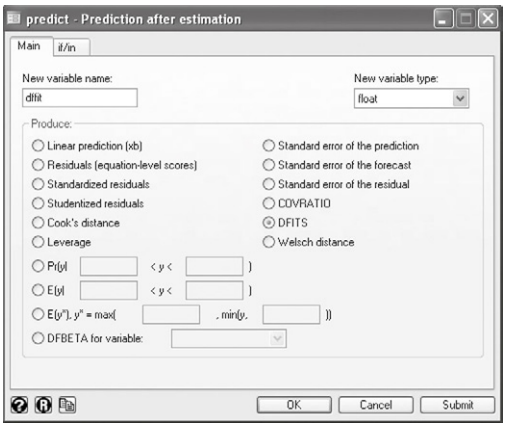


Figura 6.4 *Janela de configurações do comando* **predict** *selecionando-se a opção* **DFITS.**

A última medida que apresentaremos é o índice de covariância (COVRATIO). Esse índice mensura o impacto de uma observação nos erros-padrão. O impacto é considerado alto se $|\text{COVRATIO} - 1| \geq 3 \cdot k / N$, em que k é o número de parâmetros (incluindo o intercepto) e N é o tamanho da amostra.

Na janela de comandos do Stata®, digitaremos o seguinte:

predict cov, covratio
display 3 * 4/79

RESULTADOS 6.10 Gerando o índice de covariância e calculando o valor crítico.

```
. predict cov, covratio

. display 3 * 4 / 79
.15189873
```

Para verificar a existência de observações cujo índice de covariância menos 1, em módulo, seja igual ou superior a 0,152, iremos utilizar o seguinte comando:

list pais mor2 nata esc1 esc2 cov if abs(cov - 1) >= 0.152

De acordo com o índice de covariância, foram identificadas oito observações que seriam possíveis *outliers*: 6 e 33 (Resultados 6.11).

Para gerar o índice de covariância, via barra de menus, devemos selecionar as seguintes opções: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Será exibida uma tela, conforme a Figura 6.5.

RESULTADOS 6.11 Detectando *outliers* utilizando o índice de covariância.

```
. list pais mor2 nata esc1 esc2 cov if abs(cov - 1) >= 0.152

     +-------------------------------------------------------+
     | pais   mor2     nata       esc1       esc2       cov |
     |-------------------------------------------------------|
  6. |    6   30.1   23.814   87.07093   61.07163   1.170002 |
 33. |   33    4.7     21.5   96.90646   98.46043   .8122523 |
     +-------------------------------------------------------+
```

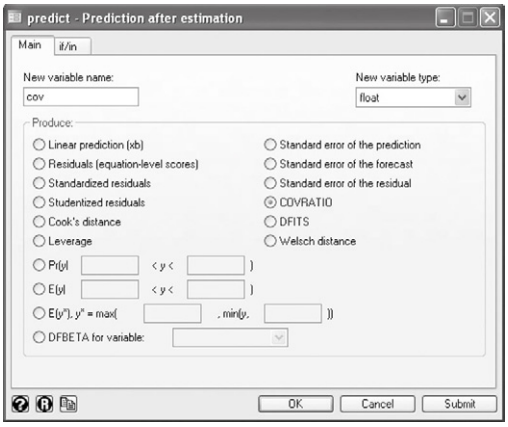


Figura 6.5 *Janela de configurações do comando* **predict** *selecionando-se a opção* **COVRATIO.**

6.2. MODELOS

Os modelos de regressão robusta visam ajustar as estimações realizadas pelo método dos mínimos quadrados, considerando-se as particularidades da amostra. Na maioria das vezes, a presença de *outliers* faz com que os pressupostos necessários para a consistência do estimador dos mínimos quadrados não sejam alcançados.

Existem três principais modelos de regressão robusta: (i) regressão com erro-padrão robusto, (ii) regressão robusta com mínimos quadrados ponderados e (iii) regressão quantílica.

Retornando ao nosso exemplo, iremos verificar se os pressupostos do estimador dos mínimos quadrados foram observados.

Na janela de comandos do Stata®, iremos informar os seguintes comandos:

estat hettest
estat imtest, white
predict res, residual

sfrancia res
estat vif

A partir dos resultados apresentados pelos testes solicitados (Resultados 6.12), verificamos que os resíduos possuem distribuição normal e não temos problemas de multicolinearidade.

RESULTADOS 6.12 Testes acessórios para a regressão linear múltipla.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of nata

         chi2(1)      =     6.62
         Prob > chi2  =   0.0101

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     15.93
         Prob > chi2  =    0.0070

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |      15.93      5    0.0070
            Skewness |       5.88      2    0.0528
            Kurtosis |       0.49      1    0.4856
---------------------+-----------------------------
               Total |      22.30      8    0.0044
---------------------------------------------------

. predict res, residual

. sfrancia res

                  Shapiro-Francia W' test for normal data

    Variable |    Obs       W'          V'        z       Prob>z
-------------+-------------------------------------------------
         res |     79    0.97769      1.674     1.003     0.15784

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
        esc2 |      4.07    0.245887
        expe |      4.07    0.245887
-------------+----------------------
    Mean VIF |      4.07
```

Todavia, em ambos os testes para a detecção de heterocedasticidade, com nível de significância de 5%, rejeitamos a hipótese nula de que os resíduos sejam homocedásticos.

A ocorrência da heterocedasticidade faz com que os parâmetros estimados estejam enviesados. Provavelmente a heterocedasticidade decorre da presença dos *outliers*, conforme vimos anteriormente.

A regressão com erro-padrão robusto permite que a estimação obtenha estimadores não enviesados. No Stata®, podemos realizar esse procedimento por meio do comando **regress**, que já estudamos, porém agora com uma nova opção (Sintaxe 6.3).

SINTAXE 6.3 Comando regress.

regress depvar indepvars [, robust] [, cluster(groupvar)]

Em que:

- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- robust: Utiliza o erro-padrão robusto à heterocedasticidade e à ausência de normalidade (estimador de Huber-White).
- cluster: Utiliza o erro-padrão robusto, porém, considerando os grupos formados a partir da variável de grupo (*groupvar*).

Para realizar uma nova estimação, iremos informar, na janela de comandos do Stata®, o seguinte:

reg nata expe esc2, robust

Na estimação utilizando o erro-padrão robusto (Resultados 6.13), verificamos que não há alteração dos coeficientes estimados. Todavia, as estatísticas utilizadas nos testes t

RESULTADOS 6.13 Resultados da regressão múltipla com erro-padrão robusto.

```
. reg nata expe esc2, robust

Linear regression                               Number of obs =      79
                                                F(  2,    76) =   95.82
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7834
                                                Root MSE      =  4.5771

------------------------------------------------------------------------------
             |               Robust
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.2912442   .1332656    -2.19   0.032    -.5566657   -.0258227
        esc2 |  -.2487269   .0467293    -5.32   0.000    -.3417965   -.1556574
       _cons |   58.06357   7.097591     8.18   0.000      43.9275    72.19965
------------------------------------------------------------------------------
```

e F são alteradas, visando corrigir os efeitos da presença de heterocedasticidade que há nos resíduos.

Após a estimação de uma regressão utilizando o erro-padrão robusto, o Stata® não permitirá a realização de testes para a detecção de homocedasticidade pois esse pressuposto não é válido para o estimador realizado.

Para realizarmos uma regressão utilizando o erro-padrão robusto, por intermédio da barra de menus, precisamos selecionar as seguintes opções: *Statistics* → *Linear models and related* → *Linear regression*. Aparecerá uma tela, conforme a Figura 6.6.

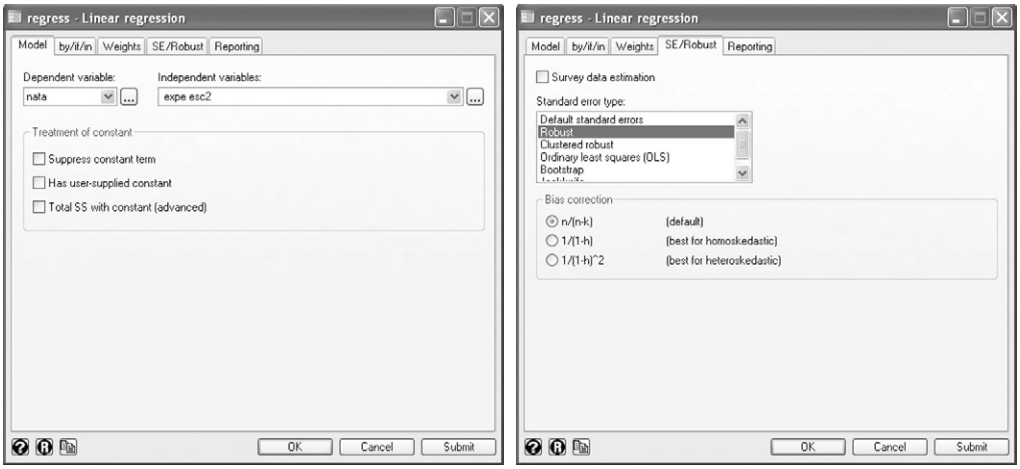


Figura 6.6 ***Janelas de configurações do comando*** **regress.**

Outra forma de se utilizar a regressão com erro-padrão no Stata® ocorre quando temos uma variável de grupo (*cluster*). Pode ocorrer que as observações que pertencem a um mesmo grupo possuam comportamento diferente quando compadadas àquelas pertencentes a outro grupo. Busca-se garantir que os resíduos das observações de um grupo não estejam correlacionados com os resíduos das demais observações nos outros grupos.

No nosso exemplo, verificamos que a variável *pop* possui uma distribuição assimétrica, indicando haver diferenças entre os países da amostra. Utilizaremos essa variável para criar uma nova variável de grupo, considerando faixas populacionais. Será elaborado o comando **gen** com a função **autocode** (Sintaxe 6.4).

SINTAXE 6.4 Comando gen com a função autocode.

gen newvar = autocode(varname, groups, min, max)

Em que:

- newvar: Variável de grupo a ser criada.
- varname: Variável quantitativa a ser utilizada para a criação de faixas.
- groups: Quantidade de grupos a serem criados.
- min: Valor mínimo a ser observado, na criação dos grupos.
- max: Valor máximo a ser observado, na criação dos grupos.

Assim sendo, precisaremos saber quais os valores mínimo e máximo da variável *pop*. Digitaremos o seguinte comando:

sum pop

Conhecendo os valores limites da variável (Resultados 6.14), iremos solicitar a criação de 15 faixas, como também verificar a quantidade de grupos formados. Para tanto, digitaremos os seguintes comandos:

RESULTADOS 6.14 Obtendo os valores mínimo e máximo da variável *pop*.

```
. sum pop

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pop |         79    2.78e+07    5.09e+07      109269   3.07e+08
```

gen rpop = autocode(pop, 15, 109269, 3.07e08)

tab rpop

Podemos observar que foram criados 10 grupos e que o primeiro é composto pela maioria dos países da amostra (Resultados 6.15). Após a criação da variável de grupo, passaremos à nova estimação utilizando a opção **cluster**.

reg nata expe esc2, cluster(rpop)

Na estimação utilizando o erro-padrão robusto e a opção **cluster**, verificamos que, novamente, não há alteração dos coeficientes estimados (Resultados 6.16). Todavia, as

RESULTADOS 6.15 Criando grupos a partir da variável *pop*.

```
. gen rpop = autocode(pop, 15, 109269, 3.07e08)

. tab rpop

       rpop |      Freq.     Percent        Cum.
------------+-----------------------------------
   2.06e+07 |         55       69.62       69.62
   4.10e+07 |          8       10.13       79.75
   6.15e+07 |          5        6.33       86.08
   8.19e+07 |          5        6.33       92.41
   1.02e+08 |          1        1.27       93.67
   1.23e+08 |          1        1.27       94.94
   1.43e+08 |          1        1.27       96.20
   1.84e+08 |          1        1.27       97.47
   2.46e+08 |          1        1.27       98.73
   3.07e+08 |          1        1.27      100.00
------------+-----------------------------------
      Total |         79      100.00
```

RESULTADOS 6.16 Resultados da regressão múltipla com erro-padrão robusto e opção cluster.

```
. reg nata expe esc2, cluster(rpop)

Linear regression                               Number of obs =      79
                                                F(  2,     9) =  681.72
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7834
                                                Root MSE      =  4.5771

                                  (Std. Err. adjusted for 10 clusters in rpop)
------------------------------------------------------------------------------
             |               Robust
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.2912442   .0355677    -8.19   0.000    -.3717038   -.2107846
        esc2 |  -.2487269   .0133332   -18.65   0.000    -.2788887   -.2185651
       _cons |   58.06357    2.06409    28.13   0.000     53.39428    62.73287
------------------------------------------------------------------------------
```

estatísticas utilizadas nos testes t e F são alteradas, utilizando-se os grupos contidos na variável *rpop*. De acordo com os resultados, verificamos que todas as variáveis foram consideradas significativas.

Para acessar o comando **generate** (ou simplesmente **gen**) por meio da barra de menus, será necessário clicar nas seguintes opções: *Data* → *Create or change data* → *Create new variable*. Surgirá uma janela, conforme a Figura 6.7.

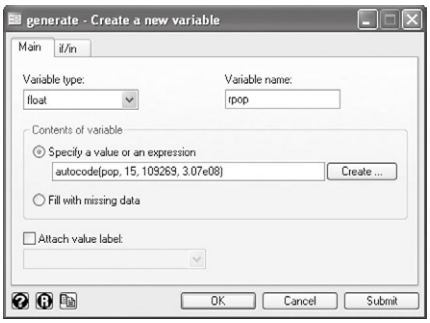


Figura 6.7 ***Janela de configurações do comando*** **gen.**

Caso quiséssemos acessar a regressão robusta com o uso da variável de grupo, via barra de menus, precisaríamos acessar as seguintes opções: *Statistics* → *Linear models and related* → *Linear regression*. Será exibida uma janela, conforme a Figura 6.8.

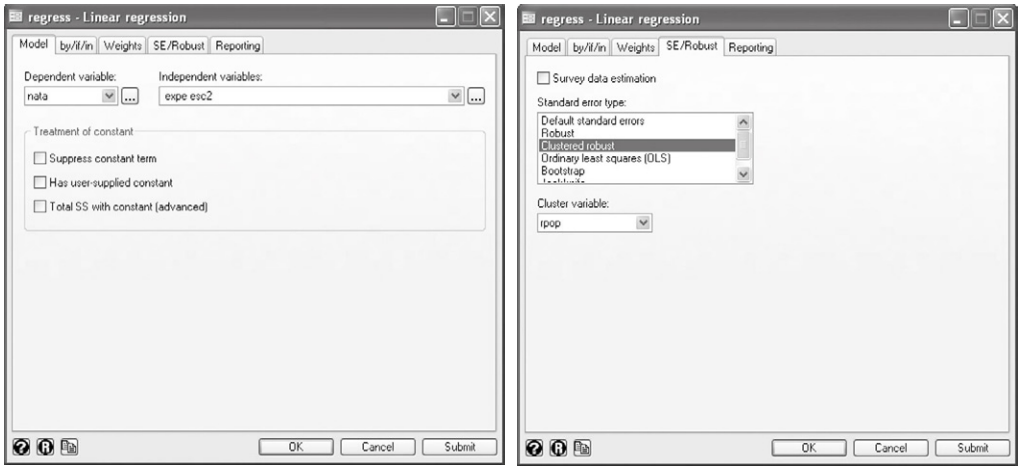


Figura 6.8 ***Janelas de configurações do comando*** **regress.**

O segundo modelo que analisaremos é a regressão robusta com mínimos quadrados ponderados. Esse modelo atribui um peso a cada observação, sendo que as observações consideradas *outliers* recebem pesos mais baixos do que as observações consideradas normais. As observações cujas distâncias de Cook forem superiores a 1 terão pesos quase nulos, de modo que não afetarão a análise do todo.

No Stata®, a regressão robusta com o estimador dos mínimos quadrados ponderados é realizada por intermédio do comando **rreg** (Sintaxe 6.5).

SINTAXE 6.5 Comando rreg.

rreg depvar indepvars [, level (#)]

Em que:

- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- level: Estabelece o nível de confiança, a ser utilizado. O padrão é 95%.

Voltando para o nosso exemplo, iremos agora realizar uma regressão robusta utilizando o comando **rreg**.

rreg nata expe esc2

Ao compararmos os resultados da regressão robusta (Resultados 6.17) com o modelo anterior, verificamos que os coeficientes estimados não são os mesmos, assim como as estatísticas dos testes t e F.

RESULTADOS 6.17 Resultados da regressão múltipla robusta.

```
. rreg nata expe esc2

   Huber iteration 1:  maximum difference in weights = .45877655
   Huber iteration 2:  maximum difference in weights = .0604747
   Huber iteration 3:  maximum difference in weights = .03721806
Biweight iteration 4:  maximum difference in weights = .1544047
Biweight iteration 5:  maximum difference in weights = .03512046
Biweight iteration 6:  maximum difference in weights = .00601326

Robust regression                                     Number of obs =      79
                                                      F(  2,    76) =  125.48
                                                      Prob > F      =  0.0000

------------------------------------------------------------------------------
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.2385164   .1148222    -2.08   0.041    -.4672047   -.0098281
        esc2 |  -.2691987   .0449875    -5.98   0.000     -.358799   -.1795984
       _cons |   55.56088   5.629818     9.87   0.000     44.34812    66.77363
------------------------------------------------------------------------------
```

Entretanto, as significâncias estatísticas dos parâmetros, bem como suas magnitudes e seus sinais, mudam muito pouco em relação ao modelo anterior.

A realização de uma regressão robusta com mínimos quadrados ponderados é possível, por meio da barra de menus, quando acessamos as seguintes opções: *Statistics* → *Linear models and related* → *Other* → *Robust regression*. Será exibida uma tela, conforme a Figura 6.9.

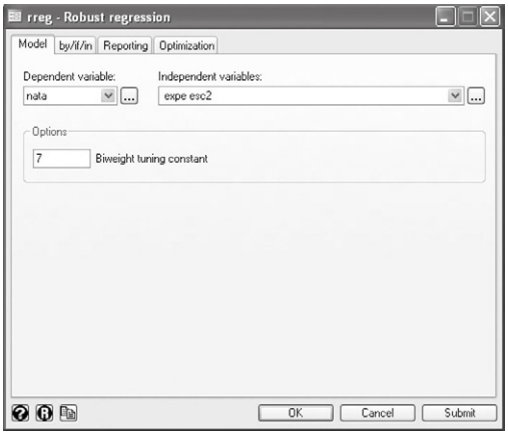


Figura 6.9 *Janela de configurações do comando* **rreg.**

O terceiro modelo de regressão robusta é a regressão quantílica, que geralmente utiliza a mediana no lugar da média, uma vez que a primeira medida de tendência é menos sensível à presença de *outliers* do que a segunda.

O comando **qreg** é utilizado no Stata® para a estimação de uma regressão quantílica (Sintaxe 6.6).

SINTAXE 6.6 Comando qreg.

qreg depvar indepvars [, level(#)] [, quantile(#)]

Em que:

- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95.
- quantile: Estabelece qual o quartil que será utilizado. O padrão é a mediana.

Voltando ao nosso exemplo, dessa vez utilizaremos a regressão quantílica para estimar os parâmetros. Digitaremos, na janela de comandos do Stata®, o seguinte:

qreg nata expe esc2

Mais uma vez, podemos notar que os coeficientes estimados são um pouco diferentes daqueles estimados pelos demais modelos (Resultados 6.18). Ocorre o mesmo em relação às estatísticas t e F. Verificamos que a variável *expe* não foi considerada significativa.

RESULTADOS 6.18 Resultados da regressão múltipla quantílica.

```
. qreg nata expe esc2
Iteration  1:  WLS sum of weighted deviations =  286.20656

Iteration  1: sum of abs. weighted deviations =  288.34847
Iteration  2: sum of abs. weighted deviations =  281.70063
Iteration  3: sum of abs. weighted deviations =  281.60575
Iteration  4: sum of abs. weighted deviations =  281.40662
Iteration  5: sum of abs. weighted deviations =  281.32964
Iteration  6: sum of abs. weighted deviations =  281.30759
Iteration  7: sum of abs. weighted deviations =  281.19304

Median regression                                   Number of obs =         79
  Raw sum of deviations  592.829 (about 17.299)
  Min sum of deviations  281.193                    Pseudo R2     =     0.5257

------------------------------------------------------------------------------
        nata |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        expe |  -.1039535    .141015    -0.74   0.463    -.3848092     .1769022
        esc2 |   -.347753   .0546021    -6.37   0.000    -.4565024    -.2390035
       _cons |   51.58224   6.890498     7.49   0.000     37.85863     65.30586
------------------------------------------------------------------------------
```

Por meio da barra de menus, podemos realizar uma regressão quantílica selecionando as seguintes opções: *Statistics* → *Nonparametric analysis* → *Quantile regression*. Será exibida uma tela, conforme a Figura 6.10.

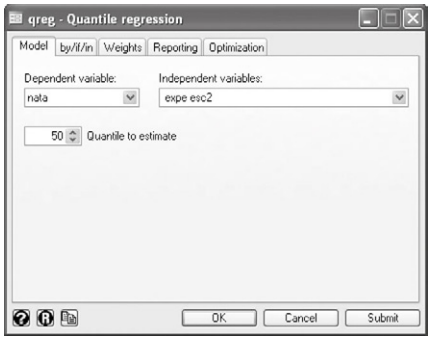


Figura 6.10 *Janela de configurações do comando* **qreg.**

6.3. EXERCÍCIO

1. O arquivo **Imovel Comercial.dta** traz dados sobre preço médio de aluguel de escritórios comerciais por metro quadrado localizados em 20 distritos municipais, bem como as taxas de vacância de cada uma destas localidades. A taxa de vacância refere-se ao percentual de área útil disponível para locação em cada distrito, calculada em relação ao estoque total do mercado em determinado período. Trata-se, portanto, de um indicador da relação entre oferta e demanda de espaços para escritórios, em dado período, induzindo ou inibindo as decisões de investimento na expansão do estoque de áreas para locação comercial e permitindo a elaboração de prognósticos envolvendo tendências de excesso de oferta no mercado.

Isto posto, pede-se:

- **a.** Elabore um gráfico de dispersão para avaliar o comportamento de *preço por metro quadrado = f (taxa de vacância)*.
- **b.** Por meio deste gráfico, é possível identificar um *outlier*?
- **c.** Elabore uma regressão linear simples não robusta a *outliers* para avaliar o comportamento de *preço por metro quadrado = f (taxa de vacância)* e salve os valores previstos gerados por meio deste modelo.
- **d.** Elabore agora uma regressão linear simples robusta a *outliers* para avaliar o comportamento de *preço por metro quadrado = f (taxa de vacância)* e salve também os valores previstos gerados por meio deste novo modelo.
- **e.** Elabore um gráfico de dispersão que contenha simultaneamente as retas correspondentes aos valores previstos em cada um dos modelos elaborados e discuta os resultados.

CAPÍTULO 7

Regressão Logística

Vamos iniciar nosso estudo da regressão logística binominal por meio da sua comparação com a regressão tradicional por mínimos quadrados ordinários. Talvez a diferença mais óbvia entre a regressão com o estimador dos mínimos quadrados ordinários e a regressão logística seja que, na primeira, a variável dependente é contínua e na regressão logística binomial, a variável dependente é uma variável codificada como 0 e 1 (*dummy*). Uma vez que a variável dependente é binária, pressupostos são mais flexíveis na regressão logística do que aqueles estabelecidos na regressão linear tradicional.

A regressão logística é similar ao método dos mínimos quadrados no sentido de se permitir identificar quais variáveis são estatisticamente significativas na análise. Diagnósticos são utilizados para avaliar se os pressupostos são válidos, havendo teste para verificar se o modelo geral é estatisticamente significativo, com um coeficiente e um erro-padrão para cada variável explicativa (UCLA, 2013).

Usaremos em nossos exemplos a base de dados **nlsw88.dta**, que comumente é instalada no mesmo diretório que o Stata®. A referida base de dados possui 2.246 observações sobre o censo norte-americano de 1988, apenas para trabalhadores do sexo feminino (Quadro 7.1).

Quadro 7.1 Variáveis que compõem a base de dados **nlsw88.dta**

Variável	Descrição	Tipo
idcode	Código	
age	Idade	Quantitativa
race	Raça (1 – branco / 2 – negro / 3 - outra)	Qualitativa
married	Estado civil (0 - solteiro / 1 - casado)	Qualitativa
never_married	Nunca casou (0 - não / 1 - sim)	Qualitativa
grade	Escolaridade em anos	Quantitativa
collgrad	Possui ensino superior (0 - não / 1 - sim)	Qualitativa
south	Mora na região sul (0 - não / 1 - sim)	Qualitativa
smsa	Mora em região metropolitana (0 - não / 1 - sim)	Qualitativa
c_city	Mora na capital (0 - não / 1 - sim)	Qualitativa
industry	Setor	Qualitativa
occupation	Ocupação	Qualitativa
union	Sindicalizado (0 - não / 1 - sim)	Qualitativa
wage	Salário por hora	Quantitativa
hours	Carga horária	Quantitativa
ttl_exp	Experiência profissional	Quantitativa
tenure	Tempo no emprego	Quantitativa

O primeiro passo será acionar o aplicativo Stata® e, após a inicialização do mesmo, iremos solicitar a abertura da base de dados **nlsw88.dta**, utilizando o comando **sysuse**.

sysuse nlsw88

RESULTADOS 7.1 Abertura do arquivo nlsw88.dta.

```
. sysuse nlsw88
(NLSW, 1988 extract)
```

7.1. REGRESSÃO LOGÍSTICA

Na regressão logística, temos o interesse em avaliar a probabilidade *p* de ocorrência de um determinado evento com base no comportamento de variáveis explicativas. Desta forma, sabendo-se que a chance de ocorrência de um evento é dada por $chance = \left(\frac{p}{1-p}\right)$, o modelo de regressão logística pode ser definido de acordo com o apresentado no Quadro 7.2.

Quadro 7.2 Modelo de regressão logística

$$\ln(chance) = Z = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad \text{[Equação 7.1]}$$

que, ao se desenvolver, chega-se a:

$$p = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}} \qquad \text{[Equação 7.2]}$$

Em que:
Z: conhecido por *logit*;
p: probabilidade estimada de ocorrência do evento de interesse;
x_i: são as variáveis explicativas, com $i = 1, 2, ..., k$; e
α e β_i: são os parâmetros do modelo.

Para ilustrarmos a diferença entre a regressão linear e a regressão logística, vamos ver o que acontece quando uma variável dependente binária é utilizada em uma regressão linear com o estimador dos mínimos quadrados ordinários.

Considere que estamos interessados em estabelecer as características, por meio das quais poderemos identificar a probabilidade de uma trabalhadora ser sindicalizada ou não (variável *union*). Inicialmente, consideraremos como variável explicativa apenas a variável *wage*.

Digitaremos na janela de comandos do Stata® o seguinte:

reg union wage

RESULTADOS 7.2 Resultados da regressão linear simples.

```
. reg union wage

      Source |       SS       df       MS              Number of obs =    1878
-------------+------------------------------           F(  1,  1876) =   44.23
       Model |  8.0124801     1   8.0124801            Prob > F      =  0.0000
    Residual |  339.824048  1876  .181142883           R-squared     =  0.0230
-------------+------------------------------           Adj R-squared =  0.0225
       Total |  347.836528  1877  .185315146           Root MSE      =  .42561

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wage |   .0156742   .0023567     6.65   0.000     .0110521    .0202963
       _cons |    .126892   .0203557     6.23   0.000     .0869698    .1668143
------------------------------------------------------------------------------
```

Como podemos observar, o Stata® realizou a estimação e exibe um resultado satisfatório para uma regressão simples. Apesar do R^2 baixo, os testes F e t indicam que o coeficiente da variável explicativa é significativo (Resultados 7.2). Entretanto, **este procedimento está errado!** Vamos observar o comportamento das variáveis nesta estimação. Escreveremos na janela de comandos o seguinte:

twoway (scatter union wage) (lfit union wage)

RESULTADOS 7.3 Gerando o gráfico de dispersão e a reta estimada pela regressão.

```
. twoway (scatter union wage) (lfit union wage)
```

No gráfico da Figura 7.1 estão plotados os valores previstos (denominados *Fitted values*; na legenda, a reta) para os valores observados da variável *union* (os pontos). Porém, ao analisarmos o gráfico, percebemos que a linha que representa as estimativas da regressão linear não é capaz de se ajustar de maneira satisfatória ao comportamento dos pontos observados.

Agora vamos realizar a mesma análise com a regressão logística. Para isso, utilizaremos o comando **logit** (Sintaxe 7.1).

SINTAXE 7.1 Comando logit.

logit depvar indepvars [, nocons] [, level(#)]

Em que:

- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- nocons: Opção a ser utilizada quando não se deseja a presença da constante no modelo regressivo.
- level: Estabelece o nível de confiança, a ser utilizado. O padrão é 95%.

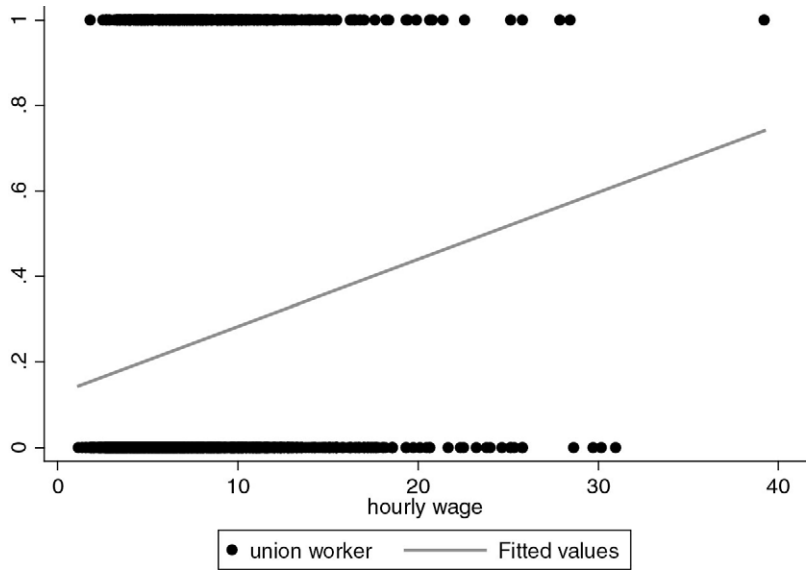


Figura 7.1 ***Gráfico de dispersão e reta estimada da regressão.***

Informaremos no Stata® o seguinte comando:
logit union wage

RESULTADOS 7.4 Resultados da regressão logística.

```
. logit union wage

Iteration 0:   log likelihood = -1046.6242
Iteration 1:   log likelihood = -1026.6546
Iteration 2:   log likelihood = -1026.3804
Iteration 3:   log likelihood = -1026.3804

Logistic regression                               Number of obs   =       1878
                                                  LR chi2(1)      =      40.49
                                                  Prob > chi2     =     0.0000
Log likelihood = -1026.3804                       Pseudo R2       =     0.0193

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wage |    .078016   .0122888     6.35   0.000     .0539304    .1021017
       _cons |  -1.737004   .1136012   -15.29   0.000    -1.959658   -1.51435
------------------------------------------------------------------------------
```

Após a estimação da regressão logística (Resultados 7.4), vamos solicitar ao Stata® que seja gerada a série de valores previstos, de acordo com o modelo estimado, para que possamos estudar a diferença entre esse modelo e o modelo de regressão linear (Figura 7.2). Utilizaremos o comando **predict** (Sintaxe 7.2).

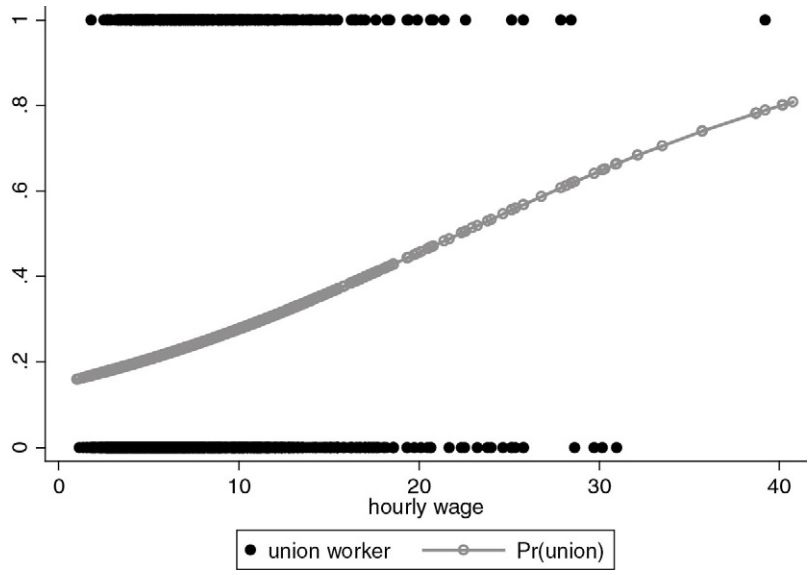


Figura 7.2 ***Gráfico de dispersão e a curva de probabilidade estimada.***

SINTAXE 7.2 Comando predict.

predict newvar [, p]

Em que:

- newvar: Nome da nova variável que armazenará os valores previstos.
- p: Opção a ser utilizada para a geração das probabilidades de acordo com o modelo da regressão.

Primeiramente, será criada a variável (*unionp*) que contém as probabilidades previstas pelo modelo para a ocorrência do evento de interesse (ser sindicalizada) para cada observação. Na sequência, os gráficos para estudo do comportamento da regressão logística serão plotados. Na janela de comandos, digitaremos o seguinte:

```
predict unionp, p
twoway (scatter union wage) (connected unionp wage, sort)
```

RESULTADOS 7.5 Gerando gráfico de dispersão e a curva de probabilidade estimada pela regressão.

```
. predict unionp, p

. twoway (scatter union wage) (connected unionp wage, sort)
```

Podemos observar que os valores estimados não formam mais uma reta, mas, sim, uma curva S. Além do mais, os valores ficam limitados entre 0 e 1. O que a regressão logística estima não são os valores da variável dependente, mas, sim, a probabilidade de ocorrência de um dos dois valores assumidos pela variável dependente (evento).

Caso desejássemos acessar o comando **logit**, utilizando a barra de menus, precisaríamos selecionar as seguintes opções: *Statistics* → *Binary outcomes* → *Logistic regression*. Surgirá uma janela, conforme a Figura 7.3.

Para acessar o comando **predict**, precisamos selecionar as seguintes opções na barra de menus: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Aparecerá uma janela, conforme a Figura 7.4.

Figura 7.3 *Janela de configurações do comando* **logit.**

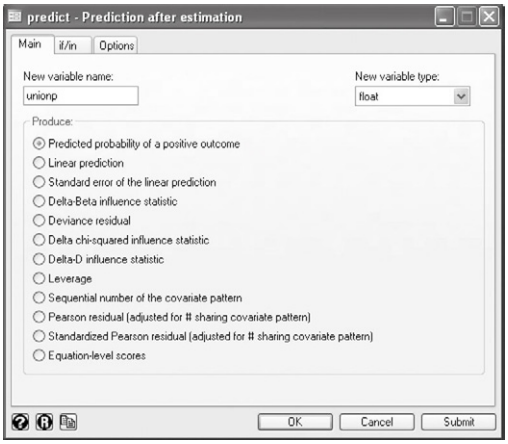


Figura 7.4 *Janela de configurações do comando* **predict.**

7.2. GRÁFICOS E ESTATÍSTICAS

Passamos agora à análise mais aprofundada da regressão logística.Vamos ampliar o nosso exemplo utilizando as seguintes variáveis explicativas: *wage*, *tenure*, *collgrad*, *south* e *c_city*.

Assim sendo, solicitamos ao Stata® que realize a seguinte regressão:

logit union wage tenure collgrad south c_city

RESULTADOS 7.6 Resultados da regressão logística.

```
. logit union wage tenure collgrad south c_city

Iteration 0:   log likelihood = -1042.6816
Iteration 1:   log likelihood = -986.93788
Iteration 2:   log likelihood = -985.79366
Iteration 3:   log likelihood = -985.79271
Iteration 4:   log likelihood = -985.79271

Logistic regression                               Number of obs   =       1868
                                                  LR chi2(5)      =     113.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -985.79271                       Pseudo R2       =     0.0546

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wage |   .0379502   .0140155     2.71   0.007     .0104803    .0654201
      tenure |   .0418208   .0096591     4.33   0.000     .0228893    .0607523
    collgrad |   .3046521   .1313169     2.32   0.020     .0472758    .5620285
       south |   -.721241   .1193708    -6.04   0.000    -.9552035   -.4872784
      c_city |   .5001731   .1181077     4.23   0.000     .2686862      .73166
       _cons |  -1.680031   .1395996   -12.03   0.000    -1.953642   -1.406421
------------------------------------------------------------------------------
```

Inicialmente, por meio da análise dos Resultados 7.6, precisamos verificar a qualidade de ajuste do modelo. De modo similar ao teste F da regressão linear, o teste da razão da verossimilhança (LR test) utiliza uma estatística com distribuição qui-quadrado para analisar a significância conjunta do modelo. As hipóteses desse teste são: H_0: todos os parâmetros são iguais a zero, e H_1: há pelo menos um parâmetro diferente de zero.

Com um p-valor inferior a 0,0001, é rejeitada a hipótese nula do teste da razão da verossimilhança e, portanto, existe pelo menos uma variável explicativa cujo parâmetro possui significância estatística no modelo logístico.

Na regressão logística, o poder explicativo do modelo é frequentemente avaliado pelo Pseudo R^2. Essa estatística é similar ao R^2 da regressão linear, porém, seu uso é mais restrito do que o R^2. O Pseudo R^2 é majoritariamente utilizado em modelos logísticos para se avaliar o ajuste quando da comparação com outros modelos.

Para verificarmos a significância individual de cada parâmetro estimado, o Stata® nos fornece o teste Z, que funciona de maneira análoga ao teste t da regressão linear. Nos resultados anteriores, verificamos que todas as variáveis explicativas e a constante foram consideradas significativas a um nível de 5%.

De acordo com os sinais estimados e o comportamento das variáveis explicativas, verificamos que, quanto maior for o salário, preservadas as demais condições, maior será a probabilidade de uma empregada ser sindicalizada. O mesmo deve ser considerado em relação ao tempo no emprego.

Em relação às *dummies collgrad* e *c_city*, notamos que, se a trabalhadora possuir nível superior e/ou morar em uma capital, aumenta a probabilidade de ser sindicalizada. Todavia, mantidas as demais condições constantes, se uma trabalhadora residir na região sul, a probabilidade de ser sindicalizada diminui.

Antes de continuarmos a análise sobre o papel de cada variável explicativa, apresentaremos outras medidas importantes para verificar o ajustamento do modelo logístico.

O teste Hosmer-Lemeshow Goodness-of-fit avalia se há diferenças significativas entre as frequências observadas e as observadas, a partir da estratificação dos valores das observações em faixas. As hipóteses do teste são as seguintes: H_0: há associação, e H_1: não há associação. Se houver associação, significa que o modelo pode ser considerado ajustado. No Stata®, a realização desse teste é feita por meio do comando **estat gof** (Sintaxe 7.3).

SINTAXE 7.3 Comando estat gof.

estat gof [, group(#)]

Em que:

- group: Caso queira que seja exibida a variável orginal do teste Hosmer-Lemeshow é necessário informar o número de grupos (#). Caso contrário, o teste será realizado com a estatística qui-quadrado de Pearson.

Devemos digitar no Stata® o seguinte comando:

estat gof

RESULTADOS 7.7 Teste Hosmer-Lemeshow.

```
. estat gof

Logistic model for union, goodness-of-fit test

       number of observations =      1868
 number of covariate patterns =      1854
           Pearson chi2(1848) =      1843.03
                 Prob > chi2 =         0.5283
```

Verificamos que, com um p-valor superior a 0,52, não rejeitamos a hipótese nula de que há associação entre os valores observados e os previstos e, consequentemente, o modelo pode ser considerado como tendo um bom ajuste (Resultados 7.7).

Para acessar o comando, via barra de menus, precisamos clicar nas seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Será exibida uma janela, conforme a Figura 7.5.

Outra forma de se avaliar um modelo logístico é observar a tabela de classificação do modelo, considerando as medidas de sensitividade, especificidade e o percentual de acerto do modelo. No Stata®, podemos solicitar a tabela de classificação do modelo por meio do comando **estat class** (Sintaxe 7.4).

SINTAXE 7.4 Comando estat class.

estat class [, cutoff(#)]

Em que:

- *cutoff*: Caso deseje alterar o ponto de corte, basta informar essa opção com o respectivo valor. Por padrão, o Stata® trabalha com um ponto de corte de 0,5.

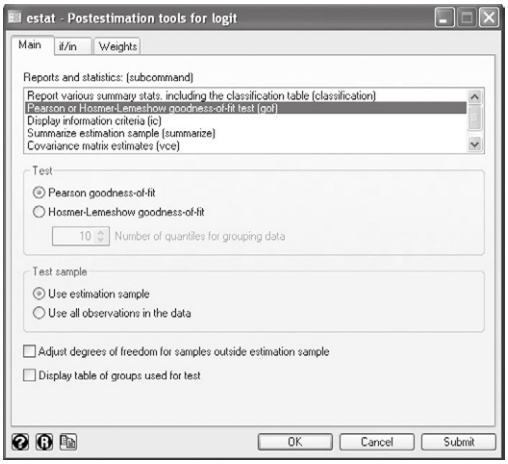


Figura 7.5 ***Janela de configurações do comando*** **estat**, ***selecionando-se a opção*** **gof.**

Solicitaremos a tabela de classificação, digitando o seguinte comando:

estat class

Na parte superior dos Resultados 7.8 são apresentados os valores observados e, na parte inferior, os valores previstos. Observamos que foram utilizadas 1.868 observações.

A sensitividade diz respeito ao total de acerto que o modelo obtém em relação ao evento (ou seja, ao fato de a trabalhadora ser sindicalizada). Podemos verificar na parte superior dos Resultados 7.8 que o modelo consegue classificar corretamente 25 trabalhadoras sindicalizadas de um total de 460 (25 / 460 = 0,0543).

A especificidade, ao contrário, se refere ao total de acertos que o modelo obtém em relação ao não evento de interesse (isto é, ao fato de a trabalhadora não ser sindicalizada). O modelo consegue classificar corretamente 1.382 trabalhadoras não sindicalizadas de um total de 1.408 (1.382 / 1.408 = 0,9815).

RESULTADOS 7.8 Tabela de classificação do modelo.

```
. estat class

Logistic model for union

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        25             26  |         51
     -     |       435           1382  |       1817
-----------+--------------------------+-----------
   Total   |       460           1408  |       1868

Classified + if predicted Pr(D) >= .5
True D defined as union != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    5.43%
Specificity                     Pr( -|~D)   98.15%
Positive predictive value       Pr( D| +)   49.02%
Negative predictive value       Pr(~D| -)   76.06%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.85%
False - rate for true D         Pr( -| D)   94.57%
False + rate for classified +   Pr(~D| +)   50.98%
False - rate for classified -   Pr( D| -)   23.94%
--------------------------------------------------
Correctly classified                        75.32%
--------------------------------------------------
```

De modo geral, o modelo logístico conseguiu classificar corretamente 75,32% das observações analisadas ([25 + 1.382] / 1.868 = 0,7532).

Para acessar o comando, por intermédio da barra de menus, precisamos clicar nas seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Será exibida uma janela, conforme a Figura 7.6.

A relação entre as estatísticas sensitividade, especificidade e ponto de corte (*cutoff*) pode ser visualizada graficamente, quando utilizamos o comando **lsens** (Sintaxe 7.5).

SINTAXE 7.5 Comando lsens.

lsens [, genp(varname1)] [, gense(varname2)] [, gensp(varname3)]

Em que:

- genp: Gera uma variável que conterá as probabilidades dos pontos de corte.
- gense: Gera uma variável que conterá a sensitividade para cada probabilidade dos pontos de corte.
- gensp: Gera uma variável que conterá a especificidade para cada probabilidade dos pontos de corte.

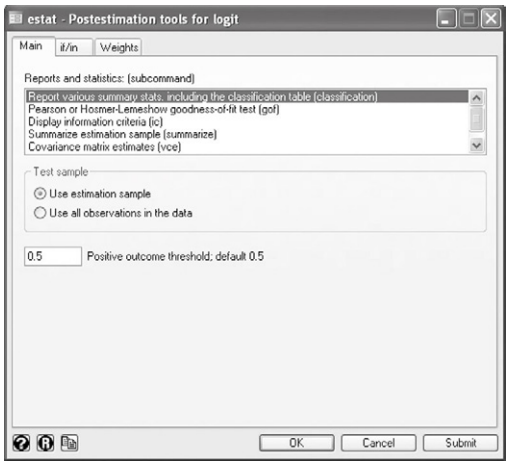


Figura 7.6 *Janela de configurações do comando* **estat,** *selecionando-se a opção* **class.**

Devemos digitar o seguinte comando:
lsens

RESULTADOS 7.9 Gerando o gráfico das probabilidades dos pontos de corte *versus* sensitividade e especificidade.

```
. lsens
```

Conforme observamos nos Resultados 7.8, o modelo com ponto de corte de 0,50 consegue prever com maior precisão as trabalhadoras não sindicalizadas do que as sindicalizadas. Se esse for o objetivo esperado do modelo, não serão necessários ajustes.

Entretanto, caso desejássemos um modelo com melhor equilíbrio entre sensitividade e especificidade, com maior sensitividade ou com mais especificidade, precisaríamos alterar o ponto de corte. A análise do gráfico apresentado na Figura 7.7 nos permitiria identificar qual seria um novo e adequado ponto de corte para o que é pretendido na análise decisória.

Vamos alterar o ponto de corte, por exemplo, para 0,25. Digitaremos na janela de comandos o seguinte:

estat class, cutoff(0.25)

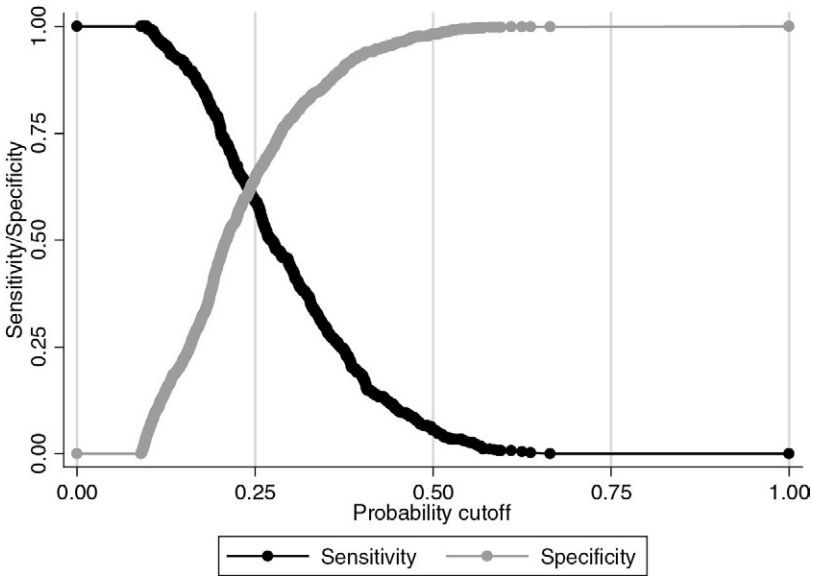


Figura 7.7 ***Gráfico das probabilidades dos pontos de corte*** **versus** ***sensitividade e especificidade.***

RESULTADOS 7.10 Tabela de classificação do modelo.

```
. estat class, cutoff(0.25)

Logistic model for union

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |       273            504  |        777
     -     |       187            904  |       1091
-----------+--------------------------+-----------
   Total   |       460           1408  |       1868

Classified + if predicted Pr(D) >= .25
True D defined as union != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   59.35%
Specificity                     Pr( -|~D)   64.20%
Positive predictive value       Pr( D| +)   35.14%
Negative predictive value       Pr(~D| -)   82.86%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   35.80%
False - rate for true D         Pr( -| D)   40.65%
False + rate for classified +   Pr(~D| +)   64.86%
False - rate for classified -   Pr( D| -)   17.14%
-------------------------------------------------
Correctly classified                        63.01%
-------------------------------------------------
```

Considerando um ponto de corte de 0,25, podemos observar que tanto o acerto geral quanto a especificidade foram menores do que na classificação anterior, que utilizou um ponto de corte de 0,50. Porém, a sensitividade, que anteriormente foi de 5,43%, passou para 59,35% (Resultados 7.10). A alteração do ponto de corte dependerá do uso que se fará do modelo regressivo e do que é pretendido pelo pesquisador em termos preditivos para uma melhor tomada de decisão.

Para acessar o comando **lsens**, por intermédio da barra de menus, precisamos clicar nas seguintes opções: *Statistics* → *Binary outcomes* → *Postestimation* → *Sensitivity/specificity plot*. Surgirá uma janela, conforme a Figura 7.8.

A curva ROC (*Receiver Operating Characteristic*) é uma medida sobre a capacidade de o modelo discriminar as categorias da variável dependente. Caso a área sob a curva seja menor ou igual a 0,5, o modelo não consegue discriminar as categorias. Se a área alcançar valores acima de 0,8, o modelo possui poder discriminatório excelente, enquanto, nos demais casos, o poder discriminatório é apenas aceitável.

No Stata®, para gerar a curva ROC (Figura 7.9), utilizamos o comando **lroc** (Sintaxe 7.6).

SINTAXE 7.6 Comando lroc.

lroc [, nograph]

Em que:

- nograph: Exibe apenas a área da curva ROC, sem gerar o gráfico.

A área sob a curva ROC é de 0,662, o que indica que o modelo não apresenta um poder discriminatório elevado (Resultados 7.11 e Figura 7.9). Percebemos essa situação quando verificamos que a sensitividade do modelo é baixa. Além disso, o Pseudo R^2 demonstra que o poder explicativo do modelo também é baixo.

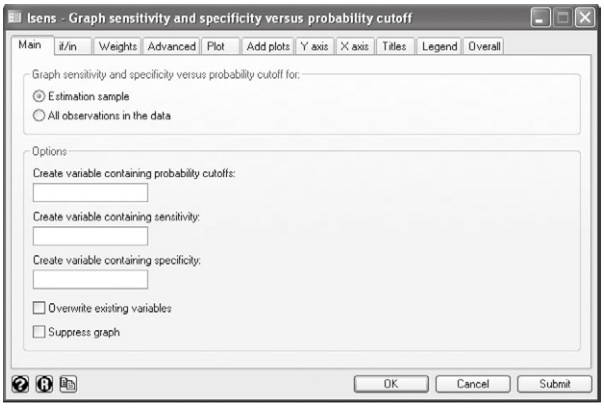


Figura 7.8 ***Janela de configurações do comando*** **lens.**

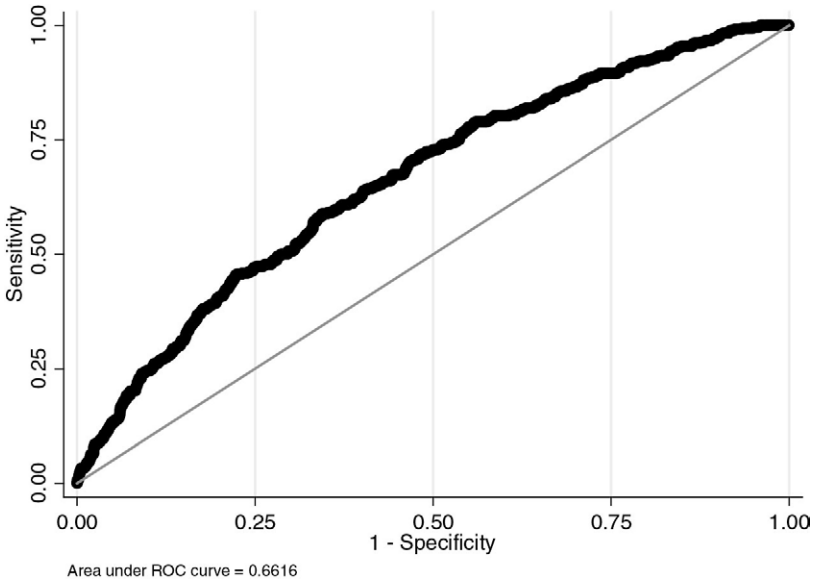


Figura 7.9 ***Curva ROC.***

RESULTADOS 7.11 Gerando a curva ROC.

```
. lroc

Logistic model for union

number of observations =      1868
area under ROC curve   =    0.6617
```

Para acessar o comando **lroc**, por meio da barra de menus, basta clicarmos nas seguintes opções: *Statistics* → *Binary outcomes* → *Postestimation* → *ROC curve after logistic/logit/probit/ivprobit*. Aparecerá uma janela, conforme a Figura 7.10.

Voltamos à análise sobre o papel de cada variável explicativa. Para isso, analisaremos o impacto dessas variáveis considerando os respectivos efeitos em relação à probabilidade de uma trabalhadora ser sindicalizada.

Para identificarmos a influência do parâmetro de cada variável explicativa sobre o comportamento da variável dependente em termos da razão de chance de ocorrência do evento em questão, ou seja, em termos de *odds ratio*, utilizaremos, no Stata®, o comando **logistic** (Sintaxe 7.7).

SINTAXE 7.7 Comando logistic.

logistic depvar indepvars [, nocons] [, level(#)]

Em que:

- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- nocons: Opção a ser utilizada quando não se deseja a presença da constante no modelo regressivo.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95%.

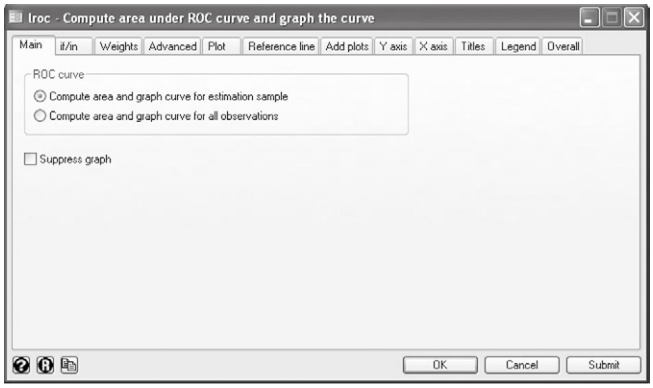


Figura 7.10 *Janela de configurações do comando* **lroc.**

Vamos, portanto, digitar o seguinte comando:

logistic union wage tenure collgrad south c_city

Os resultados apresentados são os mesmos dos obtidos com o comando **logit**, com exceção dos coeficientes estimados que não são exibidos. Ao invés destes, são apresentadas as razões de chance, ou *odds ratios* (Resultados 7.12). A razão de chance de uma variável nos informará a mudança na chance de ocorrência do evento de interesse ao se alterar em uma unidade esta mesma variável, mantidas as demais condições constantes.

Por exemplo, a cada aumento de uma unidade no salário, aumenta-se em 1,0387 vezes (um aumento de 3,87%) a chance de uma trabalhadora ser sindicalizada (1,0387 - 1 = 0,0387), mantidas as demais condições constantes. Se determinada trabalhadora morar na região sul, multiplica-se por 0,4861 vezes (uma redução de 51,39%) a chance de ser sindicalizada (0,4861 - 1 = -0,5139), mantidas as demais condições constantes. Se outra trabalhadora morar em uma capital, aumenta-se em 1,6490 vezes (um aumento de 64,90%) a chance de ser sindicalizada (1,6490 - 1 = 0,6490), também mantidas as demais condições constantes.

Por intermédio da barra de menus, podemos acessar o comando **logistic** (Figura 7.11), selecionando as seguintes opções: *Statistics* → *Binary outcomes* → *Logistic regression (reporting odds ratios)*.

RESULTADOS 7.12 Resultados da regressão logística – *odds ratio*.

```
. logistic union wage tenure collgrad south c_city

Logistic regression                                     Number of obs   =       1868
                                                        LR chi2(5)      =     113.78
                                                        Prob > chi2     =     0.0000
Log likelihood = -985.79271                             Pseudo R2       =     0.0546

------------------------------------------------------------------------------
       union | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wage |   1.038679   .0145576     2.71   0.007     1.010535    1.067607
      tenure |   1.042708   .0100716     4.33   0.000     1.023153    1.062636
    collgrad |   1.356153   .1780858     2.32   0.020     1.048411    1.754227
       south |   .4861486    .058032    -6.04   0.000     .3847338     .614296
      c_city |   1.649007   .1947605     4.23   0.000     1.308245    2.078528
       _cons |   .1863681   .0260169   -12.03   0.000     .1417569    .2450186
------------------------------------------------------------------------------
```

Figura 7.11 ***Janela de configurações do comando*** **logistic.**

7.3. REGRESSÃO LOGÍSTICA MULTINOMIAL

A regressão logística multinomial compreende uma extensão do modelo de regressão logística que permite o uso de variáveis dependentes que assumam mais de duas categorias.

Caso a variável dependente seja nominal, ou seja, não exista ordem entre suas categorias (por exemplo, candidatos em uma eleição), a regressão logística multinomial é o modelo adequado. Todavia, caso a variável dependente seja ordinal, isto é, existe uma ordem entre suas categorias (por exemplo, grande, médio e pequeno), pode ser utilizado o modelo multinomial, porém, é aconselhável o uso da regressão logística ordinal.

Suponha que agora estejamos interessados em identificar as características das trabalhadoras, considerando o setor em que atuam. Para conhecer melhor a variável *industry*

iremos inspecionar as suas categorias, utilizando o comando **tabulate** (ou simplesmente **tab**). Precisamos informar no Stata® o seguinte:

tab industry

O comando **tab** gera uma tabela de frequências para uma variável, conforme já vimos no Capítulo 2. A primeira categoria, *Ag/Forestry/Fisheries* (agricultura, extrativismo florestal e pesca), é aquela que foi codificada na entrada dos dados com o valor 1, e a última categoria, *Public Administration* (administração pública), foi codificada com o valor 12 (Resultados 7.13).

RESULTADOS 7.13 Tabela de frequências das categorias da variável *industry*.

```
. tab industry

                industry |      Freq.     Percent        Cum.
-------------------------+-----------------------------------
  Ag/Forestry/Fisheries |         17        0.76        0.76
                 Mining |          4        0.18        0.94
           Construction |         29        1.30        2.24
          Manufacturing |        367       16.44       18.68
 Transport/Comm/Utility |         90        4.03       22.72
 Wholesale/Retail Trade |        333       14.92       37.63
Finance/Ins/Real Estate |        192        8.60       46.24
    Business/Repair Svc |         86        3.85       50.09
      Personal Services |         97        4.35       54.44
  Entertainment/Rec Svc |         17        0.76       55.20
  Professional Services |        824       36.92       92.11
  Public Administration |        176        7.89      100.00
-------------------------+-----------------------------------
                  Total |      2,232      100.00
```

Para realizar a regressão logística multinomial no Stata®, faremos uso do comando **mlogit** (Sintaxe 7.8).

SINTAXE 7.8 Comando *mlogit*.

mlogit depvar indepvars [, level(#)] [, b(#)] [, rrr]

Em que:

- depvar: Nome da variável dependente.
- indepvars: Lista de variáveis explicativas.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95%.
- b: Permite identificar qual categoria será considerada como grupo de referência. Se nada for informado, o Stata® considerará a categoria da primeira observação.
- rrr: Exibe os *relative risk ratios* em vez dos coeficientes da regressão.

Na janela de comandos do Stata®, iremos informar o seguinte comando:

mlogit industry wage grade married, b(2)

O resultado do teste da razão da verossimilhança implicou um p-valor inferior a 0,0001. Logo, podemos concluir que há pelo menos uma variável estatisticamente significativa para explicar o comportamento da variável dependente, com nível de significância padrão de 5%. O Pseudo R^2 de 6,60% indica baixo poder explicativo do modelo (Resultados 7.14).

RESULTADOS 7.14 Resultados da regressão logística multinomial.

```
. mlogit industry wage grade married, b(2)

Iteration 0:   log likelihood = -4222.7456
Iteration 1:   log likelihood = -4026.1065
Iteration 2:   log likelihood = -3958.1849
Iteration 3:   log likelihood =  -3946.881
Iteration 4:   log likelihood = -3943.9995
Iteration 5:   log likelihood = -3943.9513
Iteration 6:   log likelihood = -3943.9512

Multinomial logistic regression                 Number of obs   =       2230
                                                LR chi2(33)     =     557.59
                                                Prob > chi2     =     0.0000
Log likelihood = -3943.9512                     Pseudo R2       =     0.0660

------------------------------------------------------------------------------
    industry |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Ag_Forestry_Fisheries |
        wage |  -.1995788   .0996982    -2.00   0.045    -.3949838   -.0041739
       grade |   .0693425   .2375549     0.29   0.770    -.3962565    .5349415
     married |   2.515206   1.323728     1.90   0.057    -.0792526    5.109665
       _cons |   .8258862   2.837918     0.29   0.771    -4.736331    6.388103
-------------+----------------------------------------------------------------
Mining       |  (base outcome)
-------------+----------------------------------------------------------------
Construction |
        wage |  -.0886124   .0512723    -1.73   0.084    -.1891042    .0118795
       grade |   .1226699   .2246888     0.55   0.585     -.317712    .5630518
     married |   1.342435   1.219853     1.10   0.271    -1.048432    3.733303
       _cons |   .8608366   2.692261     0.32   0.749    -4.415897     6.13757
-------------+----------------------------------------------------------------
Manufacturing |
        wage |  -.0894977   .0396788    -2.26   0.024    -.1672667   -.0117288
       grade |   .1109869   .2089112     0.53   0.595    -.2984715    .5204453
     married |   1.223724   1.164789     1.05   0.293     -1.05922    3.506669
       _cons |   3.612571   2.499414     1.45   0.148     -1.28619    8.511332
-------------+----------------------------------------------------------------
Transport_Comm_Utility |
        wage |  -.0351797   .0398423    -0.88   0.377    -.1132693    .0429098
       grade |   .2201114   .2126629     1.04   0.301    -.1967002    .6369229
     married |   1.147361   1.178119     0.97   0.330     -1.16171    3.456431
       _cons |   .3881542   2.556797     0.15   0.879    -4.623077    5.399385
-------------+----------------------------------------------------------------
Wholesale_Retail_Trade |
        wage |  -.1795196     .04327    -4.15   0.000    -.2643272   -.0947121
       grade |   .1967414   .2096558     0.94   0.348    -.2141763    .6076591
     married |   1.467627   1.166113     1.26   0.208    -.8179133    3.753167
       _cons |    2.93097   2.506819     1.17   0.242    -1.982306    7.844245
-------------+----------------------------------------------------------------
Finance_Ins_Real_Estate |
        wage |  -.0528778   .0395765    -1.34   0.182    -.1304463    .0246907
       grade |   .2460558   .2103872     1.17   0.242    -.1662955     .658407
     married |   1.878811   1.170074     1.61   0.108    -.4144923    4.172115
       _cons |   .5488874   2.524294     0.22   0.828    -4.398638    5.496413
-------------+----------------------------------------------------------------
Business_Repair_Svc |
        wage |  -.1000612   .0442188    -2.26   0.024    -.1867284   -.0133939
       grade |   .2100695   .2140772     0.98   0.326    -.2095142    .6296532
     married |   1.724749   1.182834     1.46   0.145    -.5935632    4.043062
       _cons |   .7119775   2.567922     0.28   0.782    -4.321058    5.745012
-------------+----------------------------------------------------------------
Personal_Services |
        wage |  -.4462461   .0703961    -6.34   0.000    -.5842199   -.3082723
       grade |    .108782   .2128997     0.51   0.609    -.3084937    .5260577
     married |   .9631042   1.179081     0.82   0.414    -1.347853    3.274061
       _cons |   4.379042   2.541472     1.72   0.085    -.6021515    9.360235
-------------+----------------------------------------------------------------
Entertainment_Rec_Svc |
        wage |  -.1655435   .0798219    -2.07   0.038    -.3219916   -.0090954
       grade |   .3259548   .2380307     1.37   0.171    -.1405767    .7924863
     married |   .8713547   1.258528     0.69   0.489    -1.595314    3.338024
       _cons |  -1.433333   2.868951    -0.50   0.617    -7.056374    4.189708
-------------+----------------------------------------------------------------
Professional_Services |
        wage |  -.1422114   .0400968    -3.55   0.000    -.2207997   -.0636231
       grade |   .5457693   .2090786     2.61   0.009     .1359828    .9555557
     married |   1.990576   1.163827     1.71   0.087    -.2904831    4.271636
       _cons |  -1.372782   2.504272    -0.55   0.584    -6.281066    3.535501
-------------+----------------------------------------------------------------
Public_Administration |
        wage |  -.0779153   .0402937    -1.93   0.053    -.1568896     .001059
       grade |   .3697276   .2107308     1.75   0.079    -.0432972    .7827525
     married |   1.441273   1.170401     1.23   0.218    -.8526703    3.735217
       _cons |  -.6544494   2.530015    -0.26   0.796    -5.613187    4.304288
------------------------------------------------------------------------------
```

Na regressão multinomial, as categorias são comparadas com a categoria de referência que, no nosso exemplo, se refere a *Mining* (mineração). Essa categoria foi escolhida por ser aquela com a menor quantidade de observações, porém, o critério para escolha da categoria de referência depende fundamentalmente daquilo que o pesquisador deseja.

Em relação aos testes Z, verificamos, por exemplo:

1. Entre as trabalhadoras da categoria *Professional_Services* (serviços profissionais), as variáveis *wage* e *grade* são estatisticamente significantes a um nível de 5% de significância. O mesmo ocorre com a categoria *Public Administration*, porém a um nível de significância de 10%.
2. A variável *married* somente foi significativa, a um nível de 10% de significância, para a categoria *Ag/Forestry/Fisheries*.

No modelo multinomial as razões de chances são dadas pelas *relative risk ratios*. Na janela de comandos, digitaremos:

mlogit industry wage grade married, b(2) rrr

Conforme vimos na regressão logística binária, essas chances nos permitem entender o efeito de cada variável, só que agora para cada uma das categorias analisadas (Resultados 7.15). Por exemplo, considerando a variável *wage* e um nível de significância de 5%, veremos que o efeito do aumento em uma unidade dessa variável, preservadas as demais condições, modificará a chance de uma trabalhadora pertencer respectivamente a cada uma das demais categorias, em relação à categoria *Mining*, da seguinte forma:

1. Setor *Ag/Florestry/Fisheries*: redução de 18,09%.
2. Setor *Manufacturing*: redução de 8,56%.
3. Setor *Wholesale/Retail Trade*: redução de 16,43%.
4. Setor *Business/Repair Svc*: redução de 9,52%.
5. Setor *Personal Services*: redução de 36,00%.
6. Setor *Entretainment/Rec Svc*: redução de 15,26%.
7. Setor *Professional Services*: redução de 13,26%.

Caso quiséssemos realizar uma regressão logística multinomial utilizando os comandos da barra de menus, bastaria que clicássemos nas seguintes opções: *Statistics* → *Categorical outcomes* → *Multinomial logistic regression*. Aparecerá uma janela, conforme a Figura 7.12.

Imagine que estejamos interessados em saber se dois grupos possuem coeficientes estatisticamente iguais. Neste caso, podemos utilizar o comando **test**, apresentado na Sintaxe 7.9.

Assim, na janela de comandos do Stata®, digitaremos o seguinte:

test [Entertainment_Rec_Svc]wage = [Professional_Services]wage
test [Public_Administration]grade = 1

No primeiro teste avaliamos se o valor do coeficiente estimado para o grupo *Entrertainment/Rec Svc* é igual ao coeficiente estimado para o grupo *Professional Services*, em relação à variável *wage*. Verificamos que, com um p-valor superior a 0,73, os coeficientes da variável *wage* são iguais, estatisticamente, nesses dois grupos (Resultados 7.16).

No segundo teste, o objetivo é verificar se o coeficiente da variável *grade*, estimado para o grupo *Public Administration,* é igual a 1. Com uma probabilidade inferior a 0,01, rejeitamos a hipótese nula testada (Resultados 7.16).

RESULTADOS 7.15 Resultados da regressão logística multinomial – *relative risk ratios*.

```
. mlogit industry wage grade married, b(2) rrr

Iteration 0:   log likelihood = -4222.7456
Iteration 1:   log likelihood = -4026.1065
Iteration 2:   log likelihood = -3958.1849
Iteration 3:   log likelihood =  -3946.881
Iteration 4:   log likelihood = -3943.9995
Iteration 5:   log likelihood = -3943.9513
Iteration 6:   log likelihood = -3943.9512

Multinomial logistic regression                 Number of obs   =       2230
                                                LR chi2(33)     =     557.59
                                                Prob > chi2     =     0.0000
Log likelihood = -3943.9512                     Pseudo R2       =     0.0660

------------------------------------------------------------------------------
    industry |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Ag_Forestry_Fisheries |
        wage |   .8190757   .0816604    -2.00   0.045      .673691    .9958348
       grade |   1.071803   .2546121     0.29   0.770     .6728341    1.707348
     married |   12.36916   16.37339     1.90   0.057     .9238065    165.6148
       _cons |   2.283904   6.481532     0.29   0.771     .0087708    594.7274
-------------+----------------------------------------------------------------
Mining       |  (base outcome)
-------------+----------------------------------------------------------------
Construction |
        wage |   .9152003   .0469244    -1.73   0.084     .8277003     1.01195
       grade |   1.130511   .2540132     0.55   0.585     .7278124    1.756023
     married |   3.828355    4.67003     1.10   0.271     .3504868    41.81699
       _cons |   2.365138   6.367569     0.32   0.749     .0120837    462.9275
-------------+----------------------------------------------------------------
Manufacturing |
        wage |   .9143903   .0362819    -2.26   0.024     .8459739    .9883397
       grade |    1.11738   .2334332     0.53   0.595     .7419514    1.682777
     married |   3.399826    3.96008     1.05   0.293      .346726    33.33703
       _cons |    37.0612   92.63128     1.45   0.148     .2763214    4970.779
-------------+----------------------------------------------------------------
Transport_Comm_Utility |
        wage |   .9654319   .0384651    -0.88   0.377     .8929102    1.043844
       grade |   1.246216   .2650238     1.04   0.301     .8214369    1.890654
     married |   3.149868   3.710918     0.97   0.330     .3129507    31.70362
       _cons |   1.474257   3.769377     0.15   0.879     .0098225    221.2703
-------------+----------------------------------------------------------------
Wholesale_Retail_Trade |
        wage |   .8356716   .0361595    -4.15   0.000     .7677223    .9096348
       grade |   1.217429    .255241     0.94   0.348      .807206    1.836128
     married |   4.338927   5.059681     1.26   0.208     .4413517    42.65598
       _cons |    18.7458   46.99232     1.17   0.242     .1377512    2551.011
-------------+----------------------------------------------------------------
Finance_Ins_Real_Estate |
        wage |   .9484959   .0375381    -1.34   0.182     .8777037    1.024998
       grade |   1.278971   .2690791     1.17   0.242      .846796    1.931713
     married |    6.54572    7.65898     1.61   0.108     .6606756    64.85249
       _cons |   1.731326   4.370376     0.22   0.828     .0122941    243.8159
-------------+----------------------------------------------------------------
Business_Repair_Svc |
        wage |   .9047821   .0400084    -2.26   0.024      .829669    .9866954
       grade |   1.233764   .2641208     0.98   0.326     .8109782     1.87696
     married |   5.611115   6.637019     1.46   0.145     .5523556    57.00062
       _cons |   2.038017    5.23347     0.28   0.782     .0132858    312.6275
-------------+----------------------------------------------------------------
Personal_Services |
        wage |   .6400262   .0450554    -6.34   0.000     .5575406    .7347153
       grade |   1.114919    .237366     0.51   0.609     .7345526    1.692248
     married |   2.619816   3.088977     0.82   0.414     .2597975    26.41841
       _cons |   79.76159   202.7118     1.72   0.085     .5476322    11617.12
-------------+----------------------------------------------------------------
Entertainment_Rec_Svc |
        wage |    .847433   .0676437    -2.07   0.038     .7247043    .9909458
       grade |   1.385353   .3297564     1.37   0.171      .868857    2.208882
     married |   2.390147   3.008066     0.69   0.489     .2028447    28.16342
       _cons |   .2385127   .6842812    -0.50   0.617     .0008619    66.00352
-------------+----------------------------------------------------------------
Professional_Services |
        wage |   .8674379   .0347815    -3.55   0.000     .8018773    .9383586
       grade |   1.725936   .3608561     2.61   0.009     1.145662    2.600115
     married |   7.319752   8.518927     1.71   0.087     .7479022    71.63874
       _cons |   .2534009   .6345849    -0.55   0.584     .0018714     34.3122
-------------+----------------------------------------------------------------
Public_Administration |
        wage |   .9250428   .0372734    -1.93   0.053     .8547984     1.00106
       grade |    1.44734   .3049992     1.75   0.079     .9576267    2.187485
     married |   4.226073    4.9462     1.23   0.218     .4262751    41.89711
       _cons |   .5197281   1.31492    -0.26   0.796     .0036494    74.01653
------------------------------------------------------------------------------
```

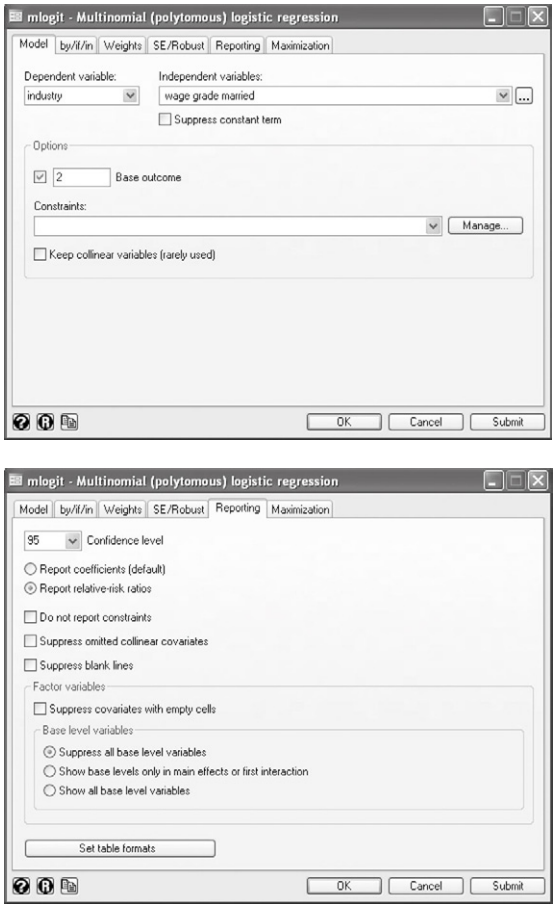


Figura 7.12 *Janelas de configurações do comando* **mlogit.**

SINTAXE 7.9 Comando **test.**

test exp

Em que:

- exp: Expressão que será considerada como hipótese nula do teste.

RESULTADOS 7.16 Resultados de testes com os coeficientes.

```
. test [Entertainment_Rec_Svc]wage = [Professional_Services]wage

 ( 1)  [Entertainment_Rec_Svc]wage - [Professional_Services]wage = 0

           chi2(  1) =    0.11
         Prob > chi2 =    0.7393

. test [Public_Administration]grade = 1

 ( 1)  [Public_Administration]grade = 1

           chi2(  1) =    8.95
         Prob > chi2 =    0.0028
```

Para realizarmos estes testes via barra de menus, basta clicar nas seguintes opções: *Statistics* → *Postestimation* → *Tests* → *Test linear hypotheses*. Aparecerá uma janela, conforme a Figura 7.13.

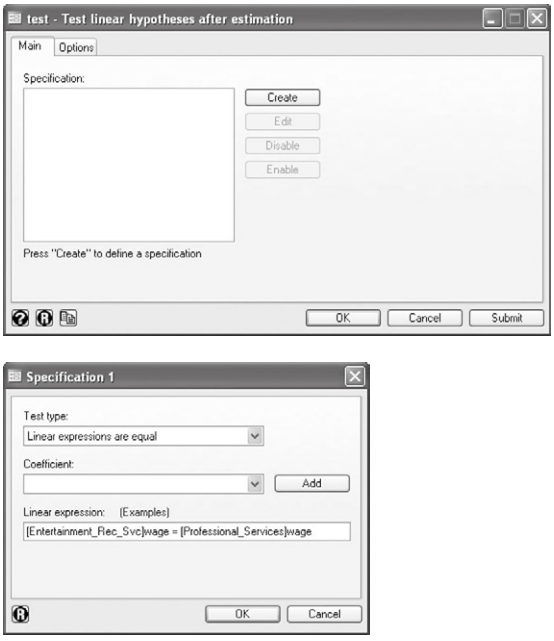


Figura 7.13 ***Janelas de configurações do comando*** **test.**

Na regressão logística podemos, ainda, observar, mediante um gráfico, o efeito de uma variável em relação às categoriais. Suponha que desejássemos conhecer qual o efeito

da variável *married* nas seguintes categorias: *Personal Services* (*industry* = 9), *Construction* (*industry* = 3) e *Public Administration* (*industry* = 12).

SINTAXE 7.10 Comando predict.

predict newvar [, outcome(groupname)] [, p]

Em que:

- newvar: Nome da nova variável que armazenará os valores previstos.
- outcome: Grupo para o qual se deseja criar os valores previstos.
- p: Opção a ser utilizada para a geração das probabilidades de acordo com o modelo da regressão.

Inicialmente, precisamos estimar as probabilidades para todas as categorias, utilizando o comando **predict** (Sintaxe 7.10).

Precisaremos informar os seguintes comandos no Stata®:

predict p01, outcome(Personal_Services) p
predict p02, outcome(Construction) p
predict p03, outcome(Public_Administration) p

RESULTADOS 7.17 Prevendo probabilidades para algumas categorias.

```
. predict p01, outcome(Personal_Services) p
(2 missing values generated)

. predict p02, outcome(Construction) p
(2 missing values generated)

. predict p03, outcome(Public_Administration) p
(2 missing values generated)
```

Após gerar as probabilidades previstas de acordo com o modelo logístico multinomial (Resultados 7.17), vamos agora plotar os gráficos confrontando essas probabilidades com a variável *wage*. Na janela de comandos, informaremos:

twoway (line p01 married if industry == 9, sort) (line p02 married if industry == 3, sort) (line p03 married if industry == 12, sort)

RESULTADOS 7.18 Gerando o gráfico para visualizar o efeito da variável *married*.

```
. twoway (line p01 married if industry == 9, sort) (line p02 married if industry == 3, sort)
>   (line p03 married if industry == 12, sort)
```

No gráfico da Figura 7.14 podemos perceber que, dentre as três categorias analisadas neste momento, o fato de a empregada ser casada tem influência apenas na probabilidade de ela pertencer à categoria *Personal Services*, com redução na chance e na probabilidade, em relação à categoria de referência (*Mining*) caso ela seja casada. Nas demais categorias analisadas, verificamos que a variável *married* não tem efeito significativo. Isso já era de se esperar, uma vez que os p-valores obtidos para esta variável nos Resultados 7.15 foram maiores do que 5% para as categorias *Construction* e *Public Administration*.

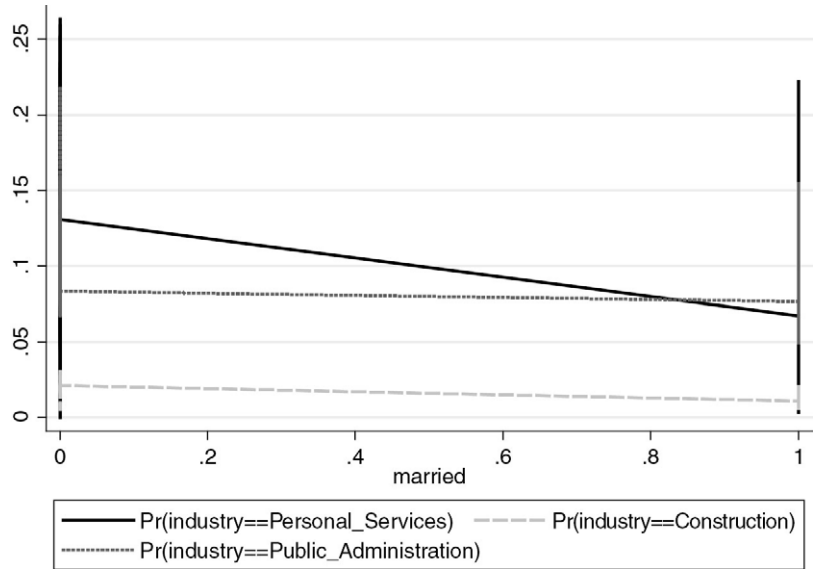


Figura 7.14 ***Gráfico para visualizar o efeito da variável*** **married.**

Para gerar as probabilidades previstas, após uma regressão logística multinomial, via barra de menus, podemos acessar as seguintes opções: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Irá aparecer uma janela, segundo a Figura 7.15.

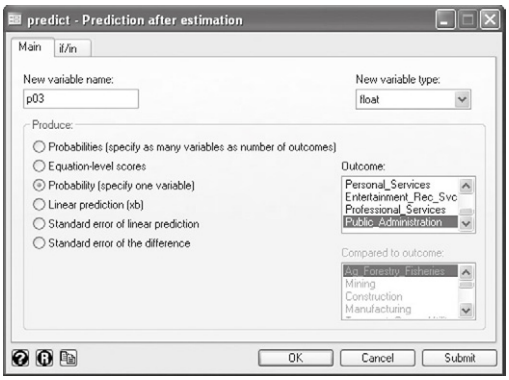


Figura 7.15 *Janela de configurações do comando* **predict.**

7.4. EXERCÍCIO

1. Para se avaliar quais as características que poderiam interferir no hábito da população em realizar exames de rotina com frequência, um pesquisador realizou uma série de entrevistas. Dados relativos a educação, idade, doenças passadas e frequência de realização de exames de rotina foram coletados.

O arquivo **medico.dta** apresenta quatro colunas (variáveis) com códigos numéricos:
Idade:

1. idade < 25
2. 25 ≤ idade ≤ 29
3. 30 ≤ idade ≤ 39
4. 40 ≤ idade ≤ 49

Educação superior (0 = Não; 1 = Sim)
Doença grave (0 = Não apresentou doença grave no passado; 1= Já apresentou alguma doença grave no passado)
Realiza exames de rotina com frequência (0 = Não; 1 = Sim)

Por meio da técnica de regressão logística, pede-se:

a. Quais variáveis são significativas para se elaborar uma boa previsão do fato de um indivíduo realizar exames de rotina com frequência?
b. Elabore novamente, sem as variáveis que apresentaram problemas de significância (teste Z).
c. Interprete os *outputs* da técnica.
d. Elabore uma curva ROC e interprete-a.
e. Calcule a probabilidade de uma pessoa com as seguintes características realizar frequentemente exames de rotina:
Idade < 25
Educação superior: Não
Doença grave no passado: Não
f. Elabore a análise de sensibilidade no Stata® e discuta os resultados.

CAPÍTULO 8

Análise de Sobrevivência: Procedimento Kaplan-Meier e Regressão de Cox

A análise de sobrevivência compreende uma variedade de métodos estatísticos destinados a analisar a duração de um evento de interesse. De acordo com Fávero *et al.* (2009), a análise de sobrevivência tem como principal vantagem o suporte a dados censurados, além de poder ser aplicada tanto nas ciências biomédicas, quanto nas ciências sociais.

Neste capítulo, apresentaremos os principais comandos relacionados com dois estimadores muito utilizados nas análises de sobrevivência: (i) Estimador de Kaplan-Meier e (ii) Regressão de Cox ou Modelo de Riscos Proporcionais.

Utilizaremos em nossos exemplos a base de dados **AIDS.dta**.[1] A referida base de dados possui 100 observações sobre tratamentos ministrados a pacientes com AIDS, sendo composto pelas variáveis contidas no Quadro 8.1.

O primeiro passo que daremos será acionar o software Stata® e, após a inicialização do mesmo, iremos solicitar a abertura da base de dados **AIDS.dta**.

8.1. DADOS CENSURADOS

Os dados utilizados em uma análise de sobrevivência apresentam duas características especiais:

1. A variável relacionada com o tempo é não negativa e, geralmente, a sua distribuição é positivamente assimétrica.
2. Para algumas observações ocorre a presença de dados censurados.

Dados censurados ocorrem quando, em algumas observações, os resultados não podem ser observados para se determinar o tempo de sobrevivência, ou porque o evento de interesse simplesmente não ocorre durante o tempo de observação ou porque há uma descontinuidade do experimento em questão (FÁVERO *et al.*, 2009).

Quadro 8.1 Variáveis que compõem a base de dados **AIDS.dta**

Variável	Descrição	Tipo
tempo de estudo	Tempo até a morte ou fim da exposição	Quantitativa
evento	1 se o paciente faleceu e 0, caso contrário	Qualitativa
remedio	Tipo de remédio	Qualitativa
idade	Idade do paciente no início da exposição	Quantitativa

[1] Banco de dados elaborado tendo por base o banco de dados **cancer.dta**, que está disponível ao se instalar o software Stata®.

Apresentaremos dois exemplos para explicar melhor o conceito de dados censurados. Conforme foi dito anteriormente, a análise de sobrevivência é um método muito utilizado nas ciências biomédicas. Imaginemos a seguinte situação: está sendo realizada uma pesquisa sobre o efeito de um medicamento e o evento analisado é a morte do paciente. Fixado o período máximo em que os pacientes serão observados, por exemplo, 180 dias, durante esse período haverá pacientes que permanecerão vivos, alguns morrerão e outros podem abandonar o tratamento. Assim, somente conheceremos o tempo de sobrevivência dos pacientes que continuarem o tratamento e que vieram a falecer durante o período.

Nas ciências sociais, podemos citar o exemplo relacionado com o risco de inadimplência de credores, pessoas físicas. Durante certo período, supondo um ano, os dados de pessoas que obtiveram empréstimos serão monitorados. O evento de interesse é a inadimplência. Assim como ocorreu no exemplo anterior, apenas conheceremos o tempo de sobrevivência dos credores que continuarem a ser monitorados e atinjam a condição de inadimplente. Se durante o período houver credores que não se tornem inadimplentes ou que deixem de ser monitorados (por exemplo, usem o benefício da portabilidade e mudem de instituição financeira), trabalharemos com dados censurados.

Quando não se considera a presença de dados censurados, a grande maioria das estimações realizadas a partir destes dados é viesada. Vamos observar o comportamento da base de dados que estamos utilizando. Na janela de comandos do Stata®, digite o seguinte comando:

sum tempo_estudo evento remedio idade

RESULTADOS 8.1 Visualizando as estatísticas descritivas das variáveis.

```
. sum tempo_estudo evento remedio idade

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
tempo_estudo |        100       16.29    10.73397         1         39
      evento |        100         .63    .4852366         0          1
     remedio |        100        1.92    .8490042         1          3
       idade |        100       55.86    5.633629        47         67
```

A variável associada ao tempo, *tempo_estudo*, possui valores mínimo e máximo de 1 e 39, respectivamente (Resultados 8.1). Esta variável é, portanto, não negativa. O evento de interesse está codificado na variável *evento* utilizando-se os valores 0 e 1. Logo, os dados são censurados. Vamos inspecionar melhor essas duas variáveis. Utilizaremos os seguintes comandos:

hist tempo_estudo

tab evento

RESULTADOS 8.2 Gerando o histograma da variável *tempo_estudo* e tabulando a variável *evento*.

```
. hist tempo_estudo
(bin=10, start=1, width=3.8)

. tab evento

     1 se o |
   paciente |
    faleceu |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         37       37.00       37.00
          1 |         63       63.00      100.00
------------+-----------------------------------
      Total |        100      100.00
```

Em relação à variável *tempo_estudo*, verificamos que a mesma é a assimétrica positivamente (Figura 8.1), conforme havíamos discutido. Ademais, de acordo com o resultado da tabulação da variável *evento*, verificamos que as observações nas quais não ocorreu o evento de interesse, durante o período analisado, compreendem 37% da nossa amostra (Resultados 8.2). Logo, estamos trabalhando com dados censurados.

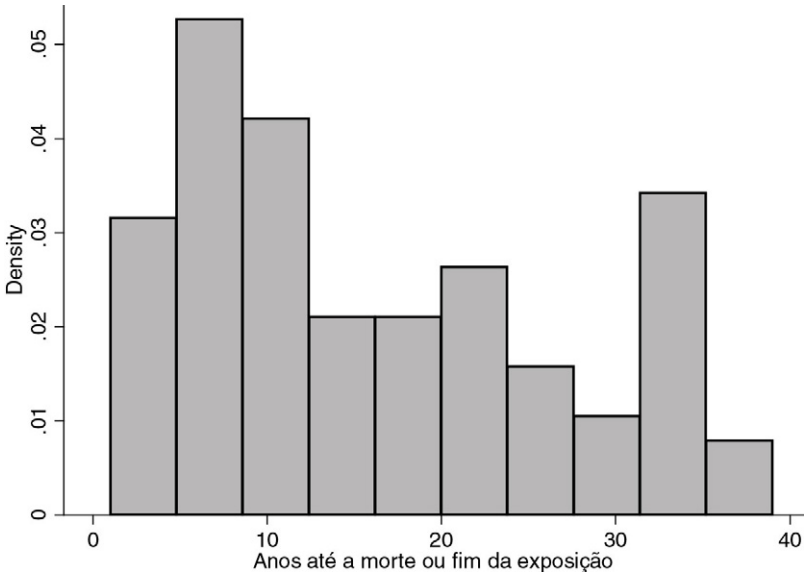


Figura 8.1 ***Histograma da variável* tempo_estudo.**

8.2. MODELOS

A análise de sobrevivência é um método que permite se examinar a duração de tempo de determinado evento. Se, por exemplo, este evento crítico for a morte, há um grande interesse pelo tempo de sobrevivência para diferentes populações por parte de empresas de seguros de vida. Por exemplo, podemos definir como tempo de sobrevivência:

- Tempo para finalização de determinado processo (empréstimo, compra de imóvel etc.) em diferentes locais ou por meio de diferentes procedimentos.
- Previsão de insolvência.
- Tempo em que diferentes grupos de consumidores manterão contas em determinado banco.

O tempo de sobrevivência pode ser considerado uma variável aleatória com distribuição de probabilidade *F(t)* e função de densidade de probabilidade *f(t)*. O interesse no uso de análise de sobrevivência é identificar a probabilidade de sobrevivência ao tempo *t*. Mais que isso, mostra-se de extremo interesse detectar a função de sobrevivência ou a curva de sobrevivência *S(t)*. A função sobrevivência, indicada por *S(t)*, pode ser definida como a probabilidade de uma observação não falhar até determinado tempo *t*, podendo ser escrita da seguinte maneira:

$$S(t) = P(T > t) = 1 - \mathrm{F}(t) \qquad \text{[Equação 8.1]}$$

$$\hat{S}(t) = \frac{N^o \text{ de observações que não falharam até momento } t}{N^o \text{ de observações no estudo}} \qquad \text{[Equação 8.2]}$$

Uma função adicional que também é de interesse na análise de sobrevivência é a função de falha ou de risco (*hazard function*), denominada por *h(t)*. Esta função representa a taxa instantânea de falha, isto é, a probabilidade de que haja a experiência de determinado evento de interesse em determinado ponto, dado que o evento ainda não ocorreu. Pode-se representar a função de falha ou de risco (*hazard function*) por:

$$h(t) = \frac{f(t)}{S(t)} \qquad \text{[Equação 8.3]}$$

$$\hat{h}(t) = \frac{N^o \text{ de observações que falharam entre } t \text{ e } t+1}{N^o \text{ de observações que não falharam até momento } t} \qquad \text{[Equação 8.4]}$$

Como explicitado pela Equação 8.3, a função de falha ou de risco apresenta o quociente entre a probabilidade instantânea de falha no período *t* e a probabilidade de sobreviver até o período *t*. Logo, a função de falha nada mais é do que uma taxa de incidência.

$$-\frac{d\log(S(t))}{dt} = h(t) \qquad \text{[Equação 8.5]}$$

E, então:

$$S(t) = \exp(-H(t)), \qquad \text{[Equação 8.6]}$$

em que *H(t)* é a função de risco integrada, também conhecida como a função de risco acumulada.

De acordo com Jenkins (2005), os modelos utilizados em uma análise de sobrevivência podem ser classificados em:

1. Modelos de riscos proporcionais (*proportional hazards models*).
2. Modelos de tempo de falha acelerado (*accelerated failure time models*).

Nos modelos de riscos proporcionais, assume-se o pressuposto de que a função de risco depende exclusivamente do tempo, e não das características das observações, ou seja, o padrão de dependência da duração é comum a todas as observações.

A interpretação dos coeficientes estimados nesses modelos relaciona a alteração de uma unidade na variável regressora a uma alteração proporcional na taxa de risco, e não no tempo de sobrevivência (JENKINS, 2005).

Nos modelos de tempo de falha acelerado, considera-se que há uma relação linear entre o logaritmo da variável temporal e as características das observações. Em razão disso, o tempo de sobrevivência pode ser curto (tempo de falha acelerado) ou longo (tempo de falha desacelerado).

A interpretação dos coeficientes estimados nos modelos de tempo de falha acelerado relaciona as alterações proporcionais em tempo de sobrevivência com a mudança em uma unidade de uma variável regressora, mantidas todas as demais fixadas (JENKINS, 2005).

8.3. ESTIMADORES

Nesta seção iremos analisar dois estimadores empregados na análise de sobrevivência: (i) Estimador de Kaplan-Meier e (ii) Regressão de Cox ou Modelo de Riscos Proporcionais.

O estimador de Kaplan-Meier é um estimador não paramétrico da função de sobrevivência. Se todas as falhas, ou períodos, em que o evento ocorre na amostra, são organizados e chamados de $t_{(j)}$ tal como $t_{(1)} \leq t_{(2)} \ldots \leq t_{(n)}$, o estimador é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right) \qquad \text{[Equação 8.7]}$$

em que d_j consiste no número de indivíduos que sofreu o evento no tempo $t_{(j)}$ e n_j é o número de indivíduos que ainda não sofreu o evento naquela ocasião e, portanto,

ainda estão expostos ao "risco" de sofrer este evento (incluídos os dados censurados em $t_{(j)}$). O produto é a apresentação de todas as falhas em um período inferior ou igual a t.

Quando a análise é realizada para apenas um grupo, examina-se a curva de sobrevivência acumulada, que apresenta as probabilidades estimadas de sobrevivência após o final de cada período. Quando muitos grupos são envolvidos, a curva de sobrevivência acumulada é elaborada para cada grupo, permitindo a comparação entre eles (teste de significância).

Iniciando o nosso exemplo, imagine que o nosso objetivo é o efeito de três tratamentos, e o evento analisado é a morte do paciente. No Stata®, primeiro precisaremos informar que a base de dados utilizada possui o formato próprio de uma análise de sobrevivência. Utilizaremos o comando **stset** (Sintaxe 8.1 e Resultados 8.3).

SINTAXE 8.1 Comando stset.

stset timevar, failure(eventvar)

Em que:

- timevar: Nome da variável relacionada com o tempo.
- eventvar: Nome da variável relacionada com o evento analisado.

Assim, digitaremos na janela de comandos do Stata® o seguinte comando:
stset tempo_estudo, failure(evento)

RESULTADOS 8.3 Definindo a amostra no formato próprio para a análise de sobrevivência.

```
. stset tempo_estudo, failure(evento)

     failure event:  evento != 0 & evento < .
obs. time interval:  (0, tempo_estudo]
 exit on or before:  failure

------------------------------------------------------------------------------
      100  total obs.
        0  exclusions
------------------------------------------------------------------------------
      100  obs. remaining, representing
       63  failures in single record/single failure data
     1629  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =        39
```

O estimador de Kaplan-Meier pode ser acessado no Stata® por meio de dois comandos: **sts** e **ltable**. Enquanto no comando **sts** (Sintaxe 8.2) a variável temporal é tratada como contínua, o comando **ltable** é indicado quando os dados da análise tiverem sido agrupados em intervalos temporais de iguais tamanhos.

SINTAXE 8.2 Comando sts.

sts [list] [graph, by(groupvar)] [gen varname = exp] [, level(#)]

Em que:

- list: Exibe as probabilidades estimadas em função do tempo de sobrevivência.
- graph: Exibe o gráfico da função de sobrevivência. Quando se usa a opção **by**, são exibidos gráficos considerando os grupos da variável *groupvar*.
- gen: Gera uma série de dados e armazena na variável *varname*, utilizando uma das seguintes expressões: **s** – função de sobrevivência, **na** – função de risco acumulada, **h** – contribuição do risco.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95%.

Vamos visualizar as probabilidades estimadas em função do tempo de sobrevivência, considerando os dados em análise. Devemos digitar o seguinte:

sts list

RESULTADOS 8.4 Probabilidades estimadas em função do tempo de sobrevivência.

```
. sts list

        failure _d:  evento
  analysis time _t:  tempo_estudo

          Beg.          Net            Survivor      Std.
  Time    Total   Fail  Lost           Function      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     1      100      4     0             0.9600     0.0196      0.8969     0.9848
     2       96      2     0             0.9400     0.0237      0.8713     0.9726
     3       94      2     0             0.9200     0.0271      0.8464     0.9592
     4       92      4     0             0.8800     0.0325      0.7984     0.9300
     5       88      4     0             0.8400     0.0367      0.7522     0.8988
     6       84      4     2             0.8000     0.0400      0.7074     0.8660
     7       78      2     0             0.7795     0.0415      0.6847     0.8489
     8       76      6     2             0.7179     0.0452      0.6182     0.7959
     9       68      0     2             0.7179     0.0452      0.6182     0.7959
    10       66      2     2             0.6962     0.0464      0.5949     0.7769
    11       62      4     2             0.6513     0.0485      0.5473     0.7371
    12       56      4     0             0.6048     0.0503      0.4988     0.6951
    13       52      2     0             0.5815     0.0510      0.4750     0.6738
    15       50      2     2             0.5582     0.0515      0.4515     0.6522
    16       46      2     0             0.5340     0.0521      0.4270     0.6296
    17       44      2     2             0.5097     0.0525      0.4029     0.6068
    19       40      0     4             0.5097     0.0525      0.4029     0.6068
    20       36      0     2             0.5097     0.0525      0.4029     0.6068
    22       34      4     0             0.4497     0.0542      0.3417     0.5518
    23       30      4     0             0.3898     0.0546      0.2833     0.4946
    24       26      2     0             0.3598     0.0544      0.2551     0.4653
    25       24      2     2             0.3298     0.0538      0.2276     0.4355
    28       20      2     2             0.2968     0.0533      0.1974     0.4028
    32       16      0     4             0.2968     0.0533      0.1974     0.4028
    33       12      3     0             0.2226     0.0545      0.1265     0.3357
    34        9      0     3             0.2226     0.0545      0.1265     0.3357
    35        6      0     3             0.2226     0.0545      0.1265     0.3357
    39        3      0     3             0.2226     0.0545      0.1265     0.3357
-------------------------------------------------------------------------------
```

A tabela resultante é composta das seguintes colunas: (i) tempo de sobrevivência (*Time*); (ii) número de indivíduos ou observações sujeitos à ocorrência do evento no tempo *t* (*Beg. Total*); (iii) número de indivíduos ou observações que sofreram o evento no tempo *t* (*Fail*); (iv) número de indivíduos ou observações que foram censurados (*Net Lost*); (v) probabilidade estimada de sobrevivência (*Survivor Function*); (vi) erro-padrão da estimação (*Std. Error*); (vii) intervalo de confiança a 95% para a probabilidade estimada de sobrevivência ao evento (*95% Conf. Int.*).

Por exemplo, quando o tempo for igual a seis anos, a probabilidade de sobrevivência é de 80%, considerando um erro-padrão de 4%. Neste exato período ocorre a primeira perda de dados (dados censurados) e, a partir de então, o denominador não será mais 100 indivíduos, já que dois indivíduos saíram da base quando *t* = 6 anos (Resultados 8.4).

Por meio da barra de menus, podemos acessar o comando **sts list**, selecionando as seguintes opções: *Statistics* → *Survival analysis* → *Summary statistics, tests, and tables* → *List survivor and cumulative hazard functions*. Surgirá uma janela, conforme a Figura 8.2.

Vamos agora gerar o gráfico da função de sobrevivência (Figura 8.3). Informaremos ao Stata® o seguinte comando:

sts graph

RESULTADOS 8.5 Gerando o gráfico da função de sobrevivência.

```
. sts graph

        failure _d:  evento
  analysis time _t:  tempo_estudo
```

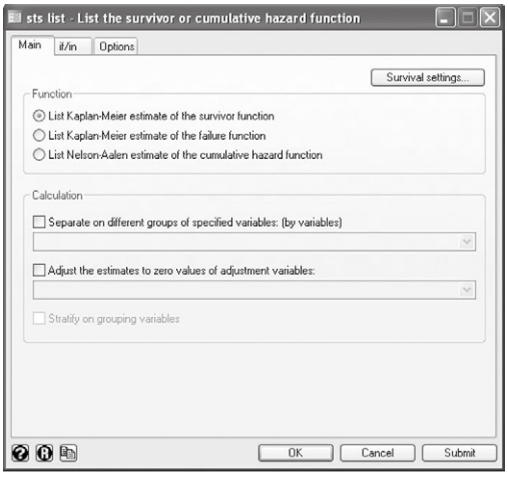


Figura 8.2 ***Janela de configurações do comando*** **sts list.**

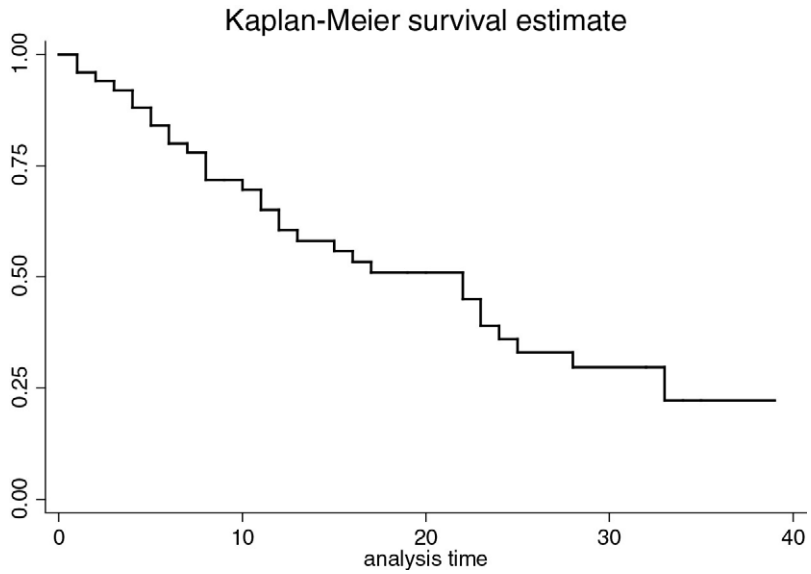


Figura 8.3 ***Gráfico da função de sobrevivência.***

Também é possível visualizar um gráfico construído a partir da função de risco acumulada (Figura 8.4) e da contribuição do risco. Precisaremos gerar as séries de cada uma destas funções, por meio do comando **sts gen**.

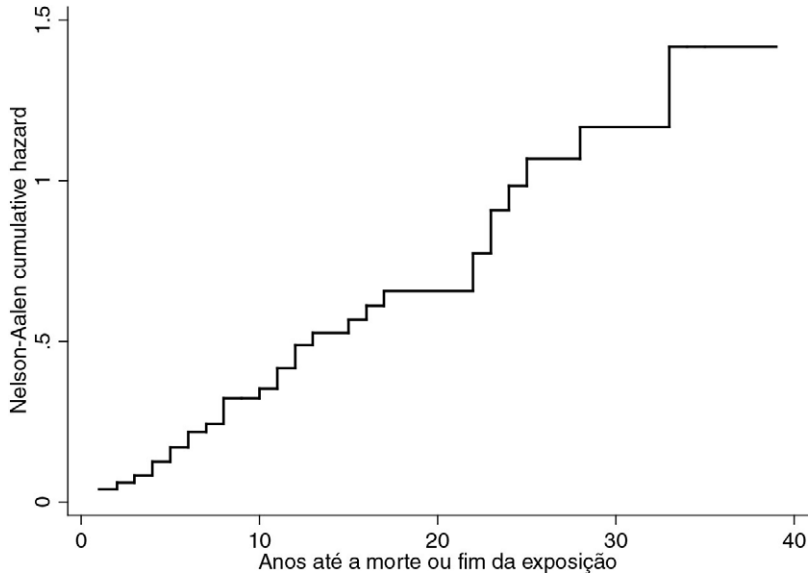


Figura 8.4 ***Gráfico da função de risco acumulada.***

Iremos solicitar a criação das séries contendo os valores calculados a partir da função de risco acumulada e da contribuição do risco. Informaremos na janela de comandos o seguinte:

sts gen ac = na
graph twoway line ac tempo_estudo, sort connect(J)

RESULTADOS 8.6 Gerando o gráfico a partir da função de risco acumulada.

```
. sts gen ac = na

. graph twoway line ac tempo_estudo, sort connect(J)
```

Por meio da observação da função de risco acumulada, podemos verificar que, de acordo com o conjunto de dados que estão sendo utilizados, à medida que o tempo avança, aumenta-se a probabilidade de ocorrência do evento de interesse. Assim, verificamos qual o comportamento do nosso evento de interesse em função do tempo. Por exemplo, poderíamos ter um evento que funcionasse em sentido contrário, ou seja, à medida que o tempo avançasse, poder-se-ia diminuir a probabilidade de ocorrência do evento.

Vamos agora analisar a contribuição do risco, para identificar momentos críticos importantes do período analisado. No Stata®, digitaremos os seguintes comandos:

sts gen ct = h
graph twoway line ct tempo_estudo, sort connect(J)

RESULTADOS 8.7 Gerando o gráfico da contribuição do risco.

```
. sts gen ct = h

. graph twoway line ct tempo_estudo, sort connect(J)
```

Com base no gráfico apresentado por meio da Figura 8.5, podemos notar as variações ocorridas no risco (contribuição do risco), calculado a partir da função de risco estimada. Por exemplo, entre o 23° e o 24° ano do período analisado, observamos que houve grandes variações, que podem ser consideradas como períodos críticos para o experimento analisado.

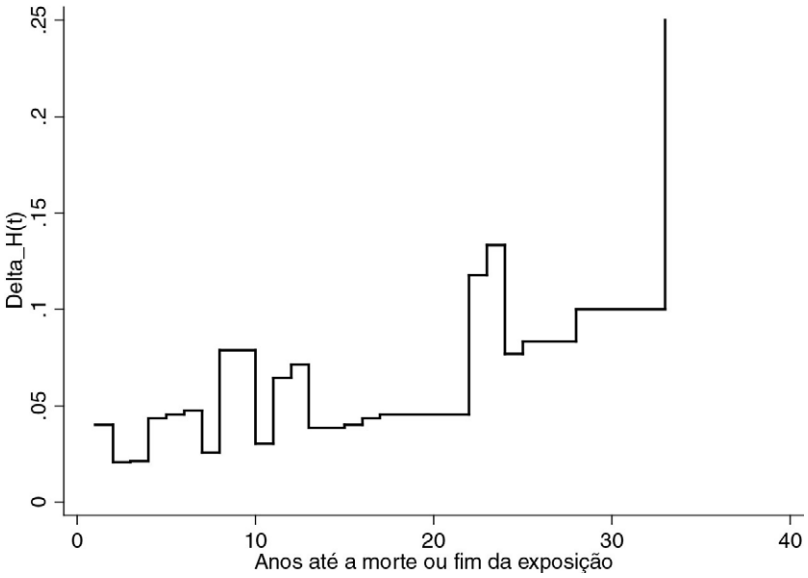


Figura 8.5 ***Gráfico da contribuição do risco.***

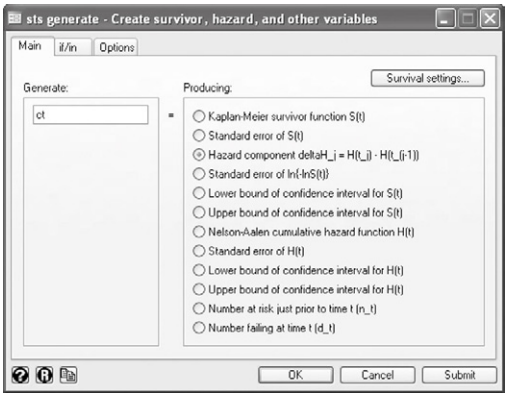


Figura 8.6 ***Janela de configurações do comando*** **sts gen.**

Para acessar este comando por meio da barra de menus, basta clicarmos nas seguintes opções: *Statistics* → *Survival analysis* → *Summary statistics, tests, and tables* → *Create survivor, hazard, and other variables*. Surgirá uma janela, conforme a Figura 8.6.

Voltaremos agora à função de sobrevivência para visualizar os efeitos de cada um dos três tratamentos utilizados. Solicitaremos a geração do gráfico da função de sobrevivência considerando o tipo de tratamento, por meio do seguinte comando:

sts graph, by(remedio)

RESULTADOS 8.8 Gerando o gráfico da função de sobrevivência por tipo de tratamento.

```
. sts graph, by(remedio)

        failure _d:  evento
  analysis time _t:  tempo_estudo
```

De acordo com a análise do gráfico da Figura 8.7, notamos que os três tipos de tratamento apresentam efeitos diferentes em relação à função de sobrevivência. Em um curtíssimo período (um ano apenas), os três tratamentos resultam na mesma probabilidade de sobrevivência. Todavia, para períodos mais longos, verifica-se que o remédio classificado como remedio = 1 mostra-se menos efetivo do que os demais tratamentos para fins de sobrevivência.

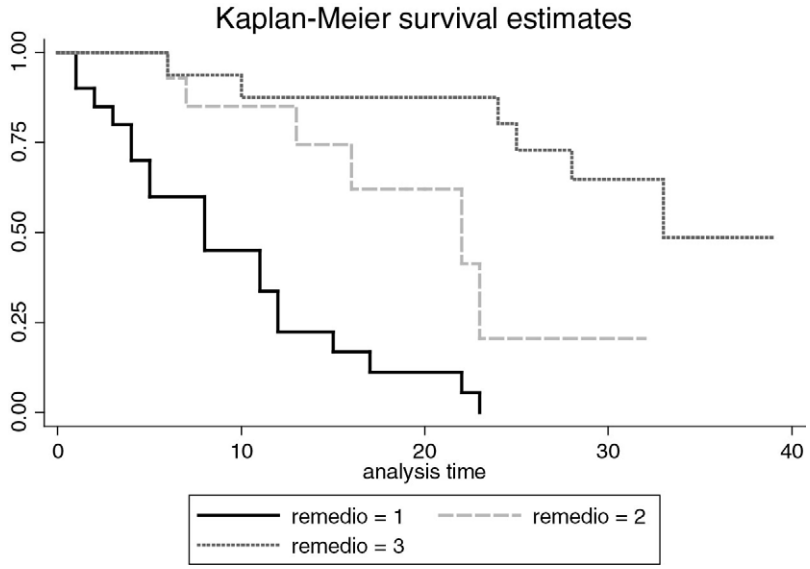


Figura 8.7 ***Gráfico da função de sobrevivência por tipo de tratamento.***

Os medicamentos classificados por remedio = 2 e remedio = 3 apresentam resultados similares até aproximadamente o oitavo ano. Entretanto, após o décimo terceiro ano, o medicamento remedio = 3 mostra-se mais efetivo contra a ocorrência do evento de interesse.

O comando **sts graph** pode ser acionado, via barra de menus. Para tanto, precisamos selecionar as seguintes opções: *Statistics* → *Survival analysis* → *Graphs* → *Survivor and cumulative hazard functions*. Irá aparecer uma janela, conforme a Figura 8.8.

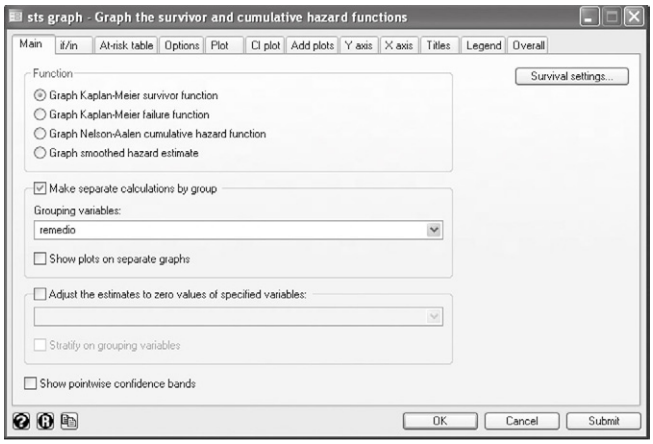


Figura 8.8 *Janela de configurações do comando* **sts graph.**

Graficamente podemos notar que há diferenças visíveis entre as funções de sobrevivência quando considerado cada um dos três tratamentos utilizados. Todavia, para atestar se as diferenças são estatisticamente significantes, precisaremos utilizar o comando **sts test** (Sintaxe 8.3).

SINTAXE 8.3 Comando sts test.

sts test varlist [if] [, w]

Em que:

- varlist: Lista de variáveis nas quais estão os grupos a serem analisados.
- if: A cláusula **if** (se) permite que o usuário estabeleça condições que limitarão a quantidade de informações que será exibida.
- w: Realiza o teste de Wilcoxon, no lugar do teste log-rank, que é a opção padrão.

Vamos, agora, verificar se há diferenças significativas entre as funções de sobrevivência, considerando-se os três tratamentos. Na janela de comandos do Stata®, digitaremos o seguinte comando:

sts test remedio

O teste long-rank verifica se há diferenças significativas a partir dos valores obtidos na função analisada, que, nesse caso, é a função de sobrevivência. As hipóteses do teste são: H_0: há igualdade entre as funções; H_1: há pelo menos uma função que é diferente (caso haja mais de duas), ou as funções são diferentes entre si (no caso de apenas duas funções).

Com um p-valor inferior a 0,0001, o teste indica a rejeição da hipótese nula (Resultados 8.9). Logo, existe pelo menos uma função que é diferente dentre as três que foram analisadas. Para realizar a comparação das funções duas a duas, precisaremos utilizar o complemento **if**, da seguinte forma:

RESULTADOS 8.9 Testando a igualdade entre as funções de sobrevivência.

```
. sts test remedio

         failure _d:  evento
   analysis time _t:  tempo_estudo


Log-rank test for equality of survivor functions

         |    Events          Events
remedio  |  observed        expected
---------+-------------------------
1        |        38           13.66
2        |        12           15.30
3        |        13           34.04
---------+-------------------------
Total    |        63           63.00

               chi2(2) =       68.76
               Pr>chi2 =      0.0000
```

sts test remedio if remedio == 1 | remedio == 2, w
sts test remedio if remedio == 1 | remedio == 3, w
sts test remedio if remedio == 2 | remedio == 3, w

O teste de Wilcoxon possui as mesmas hipóteses e finalidade do teste log-rank. Em relação aos resultados dos testes realizados (Resultados 8.10), verificamos que:

a. A função de sobrevivência do primeiro tratamento é estatisticamente diferente das funções dos outros dois tratamentos, com um nível de confiança de 99%.

b. As funções de sobrevivência do segundo e do terceiro tratamentos também são consideradas diferentes estatisticamente com um nível de significância de 5%, porém, com um nível de confiança menor do que no caso anterior (p-valor > 0,01).

Para solicitarmos a realização dos testes anteriormente apresentados, por intermédio da barra de menus, precisamos clicar nas seguintes opções: *Statistics* → *Survival analysis* → *Summary statistics, tests, and tables* → *Test equality of survivor functions*. Aparecerá uma janela, conforme a Figura 8.9.

Na sequência, iremos analisar o comando **ltable** (Sintaxe 8.4), que é indicado quando o tempo de sobrevivência, mesmo que contínuo, tenha sido observado de forma agrupada ou em valores discretos.

Para tanto, devemos observar o comportamento da função de sobrevivência e do gráfico dessa função, por intermédio do seguinte comando:

ltable tempo_estudo evento, graph

RESULTADOS 8.10 Testando a igualdade entre as funções de sobrevivência, duas a duas.

```
. sts test remedio if remedio == 1 | remedio == 2, w

         failure _d:  evento
   analysis time _t:  tempo_estudo


Wilcoxon (Breslow) test for equality of survivor functions

         |   Events          Events         Sum of
remedio  |  observed        expected         ranks
---------+-------------------------------------
1        |        38           23.10           648
2        |        12           26.90          -648
---------+-------------------------------------
Total    |        50           50.00             0

              chi2(1) =      18.85
              Pr>chi2 =     0.0000


. sts test remedio if remedio == 1 | remedio == 3, w

         failure _d:  evento
   analysis time _t:  tempo_estudo


Wilcoxon (Breslow) test for equality of survivor functions

         |   Events          Events         Sum of
remedio  |  observed        expected         ranks
---------+-------------------------------------
1        |        38           16.29          1044
3        |        13           34.71         -1044
---------+-------------------------------------
Total    |        51           51.00             0

              chi2(1) =      40.48
              Pr>chi2 =     0.0000


. sts test remedio if remedio == 2 | remedio == 3, w

         failure _d:  evento
   analysis time _t:  tempo_estudo


Wilcoxon (Breslow) test for equality of survivor functions

         |   Events          Events         Sum of
remedio  |  observed        expected         ranks
---------+-------------------------------------
2        |        12            6.09           224
3        |        13           18.91          -224
---------+-------------------------------------
Total    |        25           25.00             0

              chi2(1) =       6.10
              Pr>chi2 =     0.0135
```

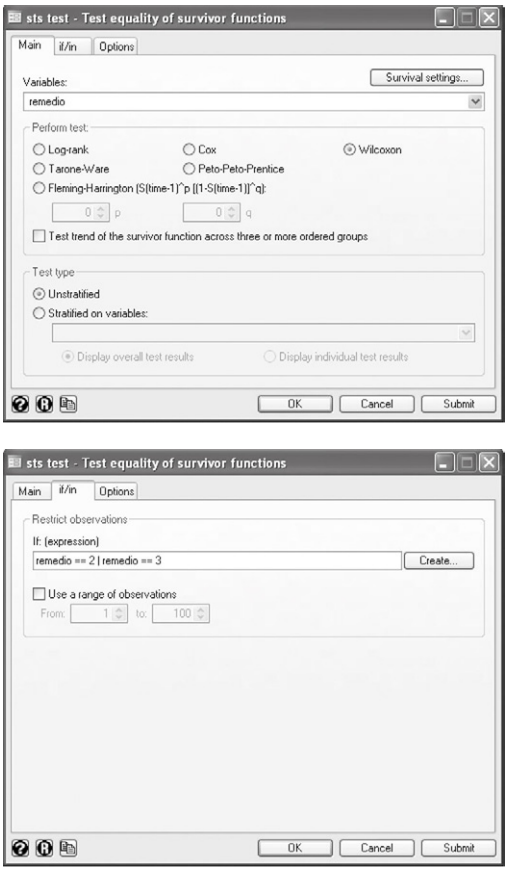


Figura 8.9 *Janelas de configurações do comando* **sts test.**

SINTAXE 8.4 Comando **ltable.**

ltable timevar deadvar [, hazard] [, failure] [, graph] [, level(#)]
[, by(groupvar)]

Em que:

- timevar: Nome da variável relacionada com o tempo.
- deadvar: Nome da variável relacionada com o evento analisado.
- hazard: Utiliza a função de risco no lugar da função de sobrevivência.
- failure: Utiliza a função de risco acumulada no lugar da função de sobrevivência.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95%.
- by: A cláusula **by** permite que o usuário separe a base em subamostras utilizando uma variável (groupvar).

A tábua de sobrevivência é similar à tabela exibida pelo comando **sts list**. Os valores

RESULTADOS 8.11 Tábua de sobrevivência.

```
. ltable tempo_estudo evento, graph

                Beg.                                    Std.
  Interval     Total   Deaths   Lost   Survival      Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     1      2     100        4      0     0.9600     0.0196     0.8969    0.9848
     2      3      96        2      0     0.9400     0.0237     0.8713    0.9726
     3      4      94        2      0     0.9200     0.0271     0.8464    0.9592
     4      5      92        4      0     0.8800     0.0325     0.7984    0.9300
     5      6      88        4      0     0.8400     0.0367     0.7522    0.8988
     6      7      84        4      2     0.7995     0.0401     0.7067    0.8657
     7      8      78        2      0     0.7790     0.0416     0.6840    0.8486
     8      9      76        6      2     0.7167     0.0454     0.6166    0.7950
     9     10      68        0      2     0.7167     0.0454     0.6166    0.7950
    10     11      66        2      2     0.6946     0.0466     0.5929    0.7757
    11     12      62        4      2     0.6491     0.0488     0.5446    0.7354
    12     13      56        4      0     0.6027     0.0505     0.4964    0.6935
    13     14      52        2      0     0.5795     0.0512     0.4728    0.6722
    15     16      50        2      2     0.5559     0.0517     0.4488    0.6503
    16     17      46        2      0     0.5317     0.0522     0.4245    0.6277
    17     18      44        2      2     0.5070     0.0526     0.3999    0.6045
    19     20      40        0      4     0.5070     0.0526     0.3999    0.6045
    20     21      36        0      2     0.5070     0.0526     0.3999    0.6045
    22     23      34        4      0     0.4473     0.0542     0.3393    0.5496
    23     24      30        4      0     0.3877     0.0546     0.2814    0.4926
    24     25      26        2      0     0.3579     0.0543     0.2534    0.4634
    25     26      24        2      2     0.3268     0.0539     0.2247    0.4326
    28     29      20        2      2     0.2924     0.0534     0.1929    0.3989
    32     33      16        0      4     0.2924     0.0534     0.1929    0.3989
    33     34      12        3      0     0.2193     0.0542     0.1239    0.3320
    34     35       9        0      3     0.2193     0.0542     0.1239    0.3320
    35     36       6        0      3     0.2193     0.0542     0.1239    0.3320
    39     40       3        0      3     0.2193     0.0542     0.1239    0.3320
-------------------------------------------------------------------------------
```

calculados apresentam, todavia, pequenas diferenças em função da forma como a variável tempo é considerada (Resultados 8.11 e Figura 8.10).

Para acessar esse comando, por intermédio da barra de menus, devemos clicar nas seguintes opções: *Statistics* → *Survival analysis* → *Summary statistics, tests, and tables* → *Life tables for survival data*. Irá surgir uma janela, conforme a Figura 8.11.

Nesse caso, podemos também comparar a sobrevivência em diferentes grupos, por meio do cálculo dos estimadores de Kaplan-Meier para a função de sobrevivência de grupos específicos e com base na aplicação de testes simples de significância (como o teste log-rank).

Entretanto, quando existir uma série de variáveis explanatórias e, em particular, quando algumas destas variáveis forem contínuas, é muito mais útil que se utilizem métodos de regressão, como a regressão de riscos proporcionais, também conhecida por regressão de Cox. Neste método, a função de risco para um indivíduo *i* é modelado como:

$$h_i(t) = h_0(t)\exp(\beta' x_i) \qquad \text{[Equação 8.8]}$$

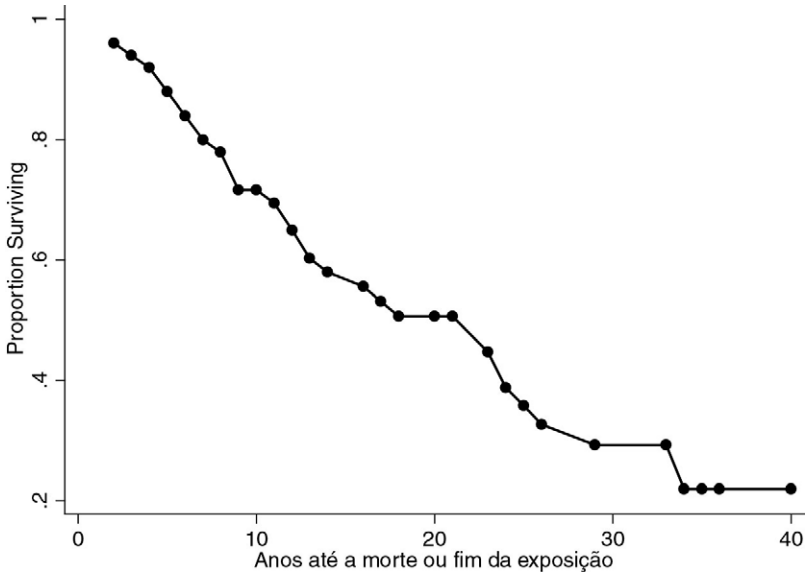


Figura 8.10 ***Gráfico da função de sobrevivência.***

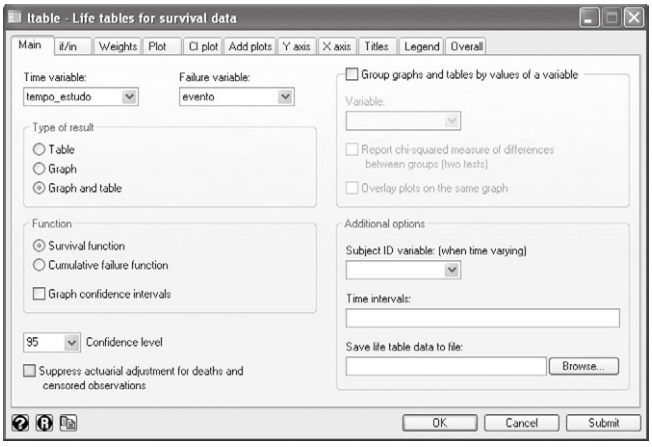


Figura 8.11 ***Janela de configurações do comando*** **ltable.**

em que β' é um vetor $px1$ de parâmetros desconhecidos, $h_0(t)$ é uma função desconhecida da taxa de falha, chamada de função de base ou basal (*baseline*), e $(\beta' x)$ é uma função conhecida, sendo usual a utilização da distribuição exponencial. Este modelo é semiparamétrico uma vez que, enquanto a função $(\beta' x)$ assume uma distribuição paramétrica, a função de base $h_0(t)$ é estimada de forma não paramétrica.

A principal suposição do modelo refere-se ao fato de que indivíduos de grupos diferentes apresentam funções de riscos proporcionais entre si, cuja razão entre as mesmas

é constante ao longo do tempo. Neste sentido, o risco de qualquer indivíduo *i* é um múltiplo da função de risco de qualquer outro indivíduo *j*, e o fator $e^{\beta'.(x_1 - x_2)}$ oferece a razão de risco (*HR*). Essa propriedade é denominada hipótese de riscos proporcionais, motivo pelo qual esta técnica também é chamada de Modelo de Riscos Proporcionais.

No Stata®, podemos realizar a regressão de Cox utilizando o comando **stcox** (Sintaxe 8.5).

SINTAXE 8.5 Comando stcox.

stcox varlist [, nohr] [, level(#)]

Em que:

- varlist: Lista de variável explicativas.
- nohr: Exibe os coeficientes e não as razões de risco, opção-padrão.
- level: Estabelece o nível de confiança a ser utilizado. O padrão é 95%.

Ainda por meio da análise de sobrevivência, iremos agora adicionar a variável *idade* e verificar o seu efeito na probabilidade de ocorrência do evento de interesse. Na janela de comandos do Stata®, digitaremos a seguinte expressão:

stcox i.remedio idade

RESULTADOS 8.12 Regressão de Cox.

```
. stcox i.remedio idade

        failure _d:  evento
  analysis time _t:  tempo_estudo

Iteration 0:   log likelihood = -251.02372
Iteration 1:   log likelihood = -213.30467
Iteration 2:   log likelihood = -211.51843
Iteration 3:   log likelihood = -211.39682
Iteration 4:   log likelihood = -211.39661
Refining estimates:
Iteration 0:   log likelihood = -211.39661

Cox regression -- Breslow method for ties

No. of subjects =          100                     Number of obs   =       100
No. of failures =           63
Time at risk    =         1629
                                                   LR chi2(3)      =     79.25
Log likelihood  =   -211.39661                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     remedio |
           2 |   .1822332   .0635977    -4.88   0.000     .0919531     .361151
           3 |   .0430619   .0201254    -6.73   0.000     .0172296    .1076244
             |
       idade |   1.116587   .0285376     4.31   0.000     1.062032    1.173944
------------------------------------------------------------------------------
```

O teste da razão da verossimilhança (*likelihood ratio*) é o mesmo utilizado na regressão logística e tem como hipóteses: H_0: todos os parâmetros são estatisticamente iguais a zero; H_1: há pelo menos um parâmetro estatisticamente diferente de zero. Com um p-valor inferior a 0,0001, verificamos que houve rejeição da hipótese nula do teste (Resultados 8.12).

Conforme já vimos no Capítulo 4, o uso do operador **i.** permite que adicionemos uma variável categórica diretamente em uma regressão. Como a variável *remedio* possui três categorias, foram criadas duas variáveis *dummies* e adicionadas ao modelo regressivo.

Individualmente, cada razão de risco (ou coeficiente, se tivesse sido utilizada a opção **nohr**) teve a sua significância estatística avaliada pelo teste Z. Verificamos que todas as variáveis explicativas do modelo foram consideradas significativas a um nível de significância de 1%.

Na regressão estimada foram apresentadas as razões de risco que funcionam de maneira similar às razões de chances (*odds ratios*) da regressão logística (Resultados 8.12). Por exemplo, quando comparamos os indivíduos que receberam o segundo tratamento com aqueles que receberam o primeiro tratamento, verificamos que a probabilidade de ocorrência do evento de interesse é reduzida em 81,78%, mantendo-se constantes as demais condições (0,1822 - 1 = -0,8178).

Quando realizamos a mesma comparação, porém, envolvendo o primeiro e o terceiro tratamentos, verificamos que a redução passa a ser de 95,7% (0,043 - 1 = – 0,957), também mantendo-se as demais condições constantes. Em relação à idade do paciente, verificamos que o aumento em uma unidade dessa variável aumenta a probabilidade de ocorrência do evento de interesse em 11,66% (1,1166 - 1 = 0,1166).

Para realizar uma estimação do modelo regressivo de Cox, utilizando a barra de menus, podemos selecionar as seguintes opções: *Statistics* → *Survival analysis* → *Regression models* → *Cox proportional hazards model*. Será exibida uma janela, conforme a Figura 8.12.

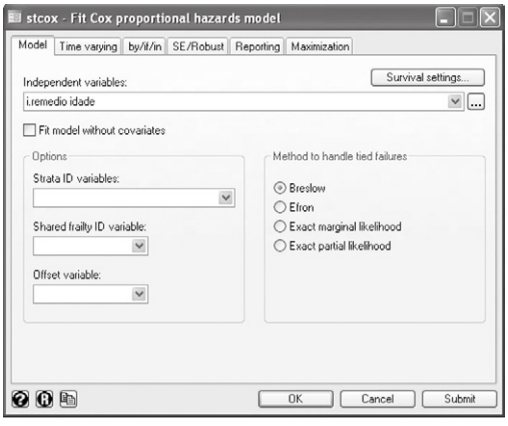


Figura 8.12 *Janela de configurações do comando* **stcox.**

Para que visualizemos os gráficos das funções de sobrevivência e de risco acumulada, assim como das contribuições do risco (Figuras 8.14 a 8.16), precisaremos gerar as respectivas séries, por meio do comando **predict** (Sintaxe 8.6).

SINTAXE 8.6 Comando predict.

predict newvar [, bases] [, basec] [, basehc]

Em que:

- newvar: Nome da nova variável que armazenará os valores previstos.
- bases: Opção a ser utilizada para a geração dos valores segundo a função de sobrevivência.
- basec: Opção a ser utilizada para a geração dos valores segundo a função de risco acumulada.
- basech: Opção a ser utilizada para a geração dos valores segundo as contribuições do risco.

Dessa forma, é necessário que solicitemos ao Stata® que sejam geradas as respectivas séries, por meio dos seguintes comandos:

```
predict cox_s, bases
predict cox_na, basec
predict cox_ct, basehc
```

RESULTADOS 8.13 Gerando as séries das funções de sobrevivência e de risco acumuladas, além das contribuições do risco.

```
. predict cox_s, bases

. predict cox_na, basec

. predict cox_ct, basehc
```

Para acessar o comando **predict**, precisamos selecionar as seguintes opções na barra de menus: *Statistics* → *Postestimation* → *Predictions, residuals, etc.* Aparecerá uma janela, conforme a Figura 8.13.

A partir das novas séries geradas, podemos solicitar a geração dos gráficos. Novamente, é importante lembrar que os comandos que geram gráficos no Stata® são exibidos na mesma janela. Então, devemos gerar e copiar (ou salvar) um gráfico, antes de solicitarmos outro. Na janela de comandos do Stata®, informaremos o seguinte:

```
twoway line cox_s tempo_estudo, sort connect(J)
twoway line cox_na tempo_estudo, sort connect(J)
```

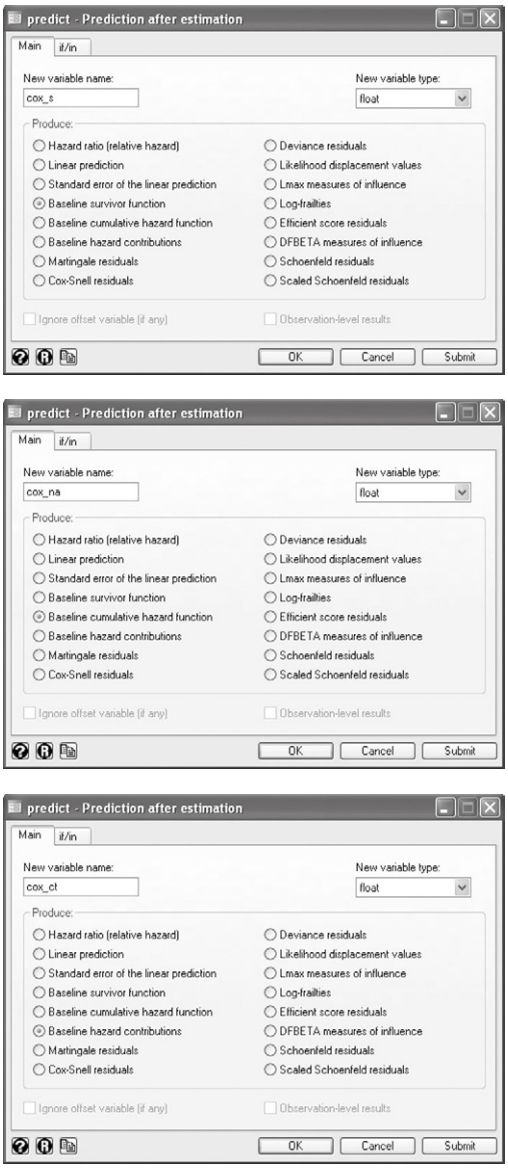


Figura 8.13 *Janelas de configurações do comando* **predict.**

twoway line cox_ct tempo_estudo, sort connect(J)

Quando comparamos estes gráficos com aqueles obtidos pelo estimador de Kaplan-Meier, verificamos que a inclusão da variável *idade* apresenta-nos uma situação bastante interessante. Até o décimo sétimo ano, a probabilidade de sobrevivência é alta. A partir desse momento, começa a haver reduções mais intensas na probabilidade de sobrevivência e, conforme vimos anteriormente, esta redução tende a ser maior quanto maior for a idade do paciente.

RESULTADOS 8.14 Gerando os gráficos da análise de sobrevivência.

```
. twoway line cox_s tempo_estudo, sort connect(J)
. twoway line cox_na tempo_estudo, sort connect(J)
. twoway line cox_ct tempo_estudo, sort connect(J)
```

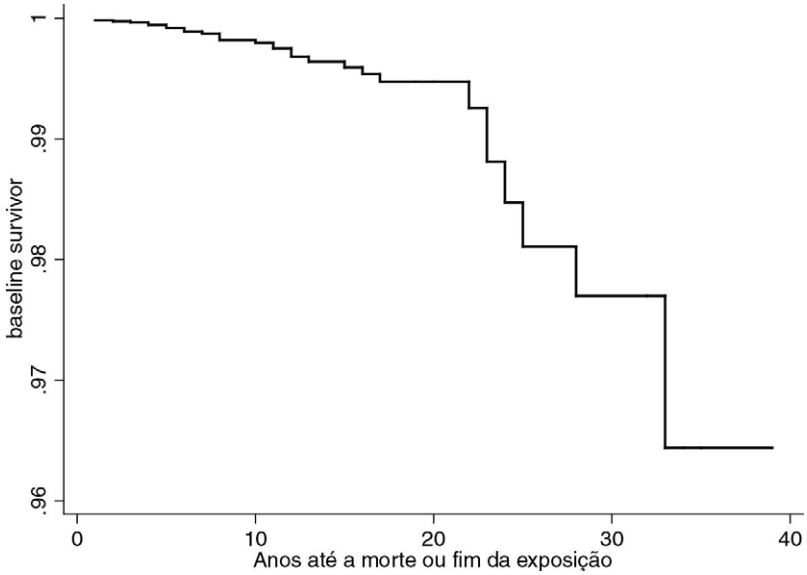


Figura 8.14 *Gráfico da função de sobrevivência.*

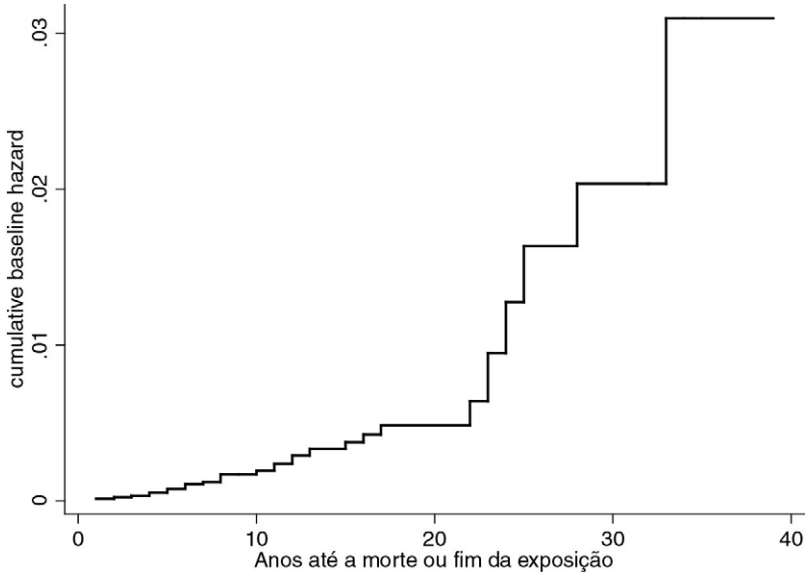


Figura 8.15 *Gráfico da função de risco acumulada.*

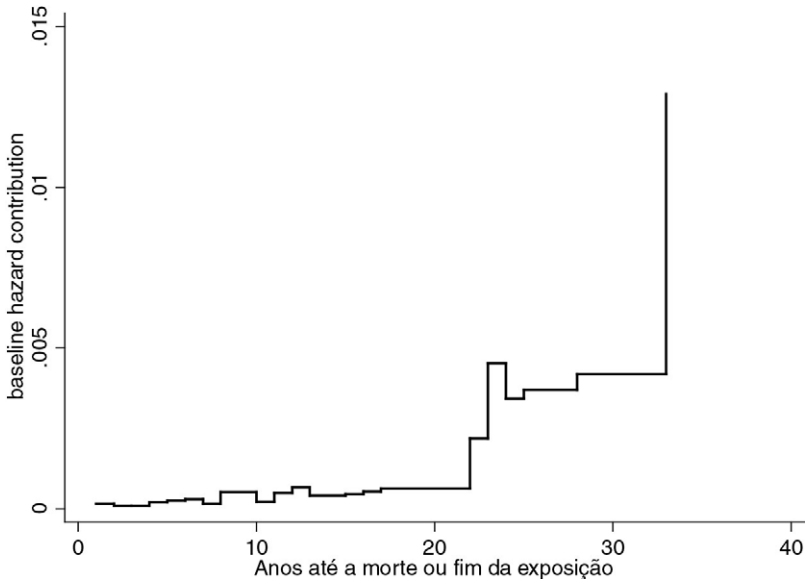


Figura 8.16 ***Gráfico das contribuições do risco.***

Quando estivermos trabalhando com mais de um modelo, podemos comparar o poder preditivo dos mesmos por meio do emprego das medidas de associação C de Harrell (Harrell's C) e D de Somers (Somers' D). Para acessá-las, utilizaremos o comando **estat concordance** (Sintaxe 8.7).

SINTAXE 8.7 Comando estat concordance.

estat concordance [, noshow]

Em que:

- noshow: Não mostra quais são as variáveis de configuração do formato utilizado na análise de sobrevivência.

Na janela de comandos do Stata®, digitaremos o seguinte:

estat concordance

As estatísticas C de Harrell e D de Somers alcançaram os valores de 0,819 e 0,638, respectivamente (Resultados 8.15). Conforme discutido, quando tivermos que escolher entre dois ou mais modelos, poderemos utilizar tais estatísticas, visto que, quanto maiores forem seus valores, maior será o poder preditivo de um modelo.

Para acessar o comando **estat concordance**, por meio da barra de menus, devemos selecionar as seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Aparecerá uma janela, conforme a Figura 8.17.

RESULTADOS 8.15 Computado o poder preditivo do modelo regressivo.

```
. estat concordance

        failure _d:  evento
  analysis time _t:  tempo_estudo

Harrell's C concordance statistic

Number of subjects (N)                =       100
Number of comparison pairs (P)        =      3651
Number of orderings as expected (E)   =      2968
Number of tied predictions (T)        =        44

         Harrell's C = (E + T/2) / P  =      .819
                          Somers' D   =     .6379
```

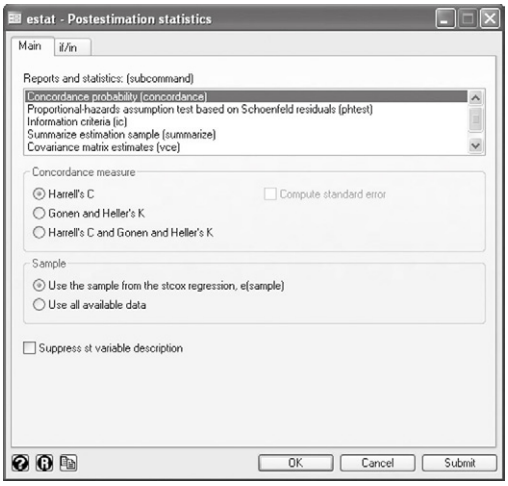


Figura 8.17 *Janela de configurações do comando* **estat,** *selecionando-se a opção* **concordance.**

Conforme discutido, a principal suposição do modelo de riscos proporcionais refere-se ao fato de que indivíduos de grupos diferentes apresentam funções de riscos proporcionais entre si, cuja razão entre as mesmas é constante ao longo do tempo. Para verificar se a amostra utilizada é realmente adequada à suposição, utilizaremos o comando **estat phtest** (Sintaxe 8.8).

Para testar se o pressuposto da proporcionalidade do risco foi observado, digitaremos na janela de comandos o seguinte:

SINTAXE 8.8 Comando estat phtest.

estat phtest [, detail]

Em que:

- detail: Além do teste geral, essa opção exibe o resultado do teste para cada regressor.

estat phtest, detail

RESULTADOS 8.16 Testando o pressuposto de proporcionalidade do risco.

```
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      1b.remedio  |         .               .        1             .
      2.remedio   |       0.13224         0.94       1          0.3311
      3.remedio   |      -0.07534         0.34       1          0.5605
      idade       |      -0.06426         0.23       1          0.6291
      ------------+---------------------------------------------------
      global test |                       1.99       3          0.5750
      ----------------------------------------------------------------
```

De acordo com os p-valores obtidos por meio do teste do pressuposto de proporcionalidade do risco, é possível verificarmos que não houve rejeição da hipótese nula de que os riscos sejam proporcionais entre si, nem no teste global, nem nos individuais para cada regressor (Resultados 8.16).

Para acessar o comando **estat phtest**, via barra de menus, devemos selecionar as seguintes opções: *Statistics* → *Postestimation* → *Reports and statistics*. Irá surgir uma janela, conforme a Figura 8.18.

8.4. EXERCÍCIOS

1. Por meio do Arquivo **AIDS.dta**, realize a análise de sobrevivência com base no procedimento *Life Table* (segregando-a segundo o tipo de droga). Sendo assim:
   - **a.** Qual a probabilidade estimada de sobrevivência dos indivíduos com AIDS após cinco anos de estudo? Demonstre os cálculos.
   - **b.** Há diferenças entre o tipo de drogas?
   - **c.** Há diferenças entre o tipo de drogas para indivíduos acima de 55 anos?
   - **d.** Há diferença na sobrevivência de indivíduos acima de 55 anos dos demais indivíduos?
2. Um pesquisador deseja modelar o tempo gasto por um estudante para obter uma pós-graduação. O arquivo **pos_graduacao.dta** contém quatro colunas:
   - Ano: codificado de 1 a 14, representando os anos desde o fim da graduação.
   - Universidade:
     - 1 para Universidade A,
     - 2 para Universidade B,

Figura 8.18 ***Janelas de configurações do comando*** **estat**, ***selecionando-se a opção*** **phtest.**

- 3 para Universidade C.
- Residência: 1 para residentes permanentes; 2 para residentes temporários.
- Evento: Número de estudantes nesta categoria.

Por intermédio do procedimento Kaplan-Meier:

a. Verifique se há diferenças entre as universidades.

b. Há diferenças entre os tipos de residência?

3. Uma estudante interessada em se casar, com o intuito de escolher o parceiro ideal, realizou uma pesquisa para determinar os principais fatores associados à sobrevivência ao evento divórcio. A unidade de observação pesquisada foram casais e o evento de interesse, o divórcio. A ausência de dados e a viuvez são tratadas como eventos censurados. As variáveis englobadas na pesquisa são, portanto:

- *id*: identificação do casal.
- *heduc*: anos de estudo do marido, codificado como:

  - 0 = menos de 12 anos,
  - 1 = 12 a 15 anos, e
  - 2 = 16 ou mais anos.
- *Cas_anterior*: codificado 1 se alguém do casal já foi casado e 0, caso contrário.
- *filhos*: codificado 1 se o casal possui filhos e 0, caso o contrário.
- *anos*: duração do casamento, desde a data do casamento até a data do divórcio ou do dado censurado.
- *div*: o indicador de falha, codificado como 1 para divórcio e 0 para dados censurados.

Por meio do procedimento Kaplan-Meier aplicado ao arquivo **divorcio.dta**:

a. Qual a probabilidade de um casal sobreviver ao divórcio depois de cinco anos de casados?

b. Verifique se há diferenças na probabilidade em se divorciar de acordo com o nível de formação dos maridos.

c. Há diferenças entre os casais com e sem filhos?

d. Há diferenças entre os casais com maridos que já foram casados e os casais com maridos que estão no primeiro casamento?

4. Ainda por meio da utilização do arquivo **divorcio.dta**, estime o Modelo de Riscos Proporcionais. Também esboce o gráfico de sobrevivência e o gráfico de risco acumulado para os casos em que o marido já é divorciado ou não.

CAPÍTULO 9

Regressão com Dados em Painel

É cada vez mais crescente e importante a utilização de modelos que envolvam dados provenientes de várias *cross-sections* ao longo do tempo (dados em painel). Como muitos dados de empresas, municípios ou países são divulgados periodicamente, o pesquisador é convidado, naturalmente, a aplicar modelos longitudinais para o estudo de fenômenos que sofrem influência das diferenças entre os indivíduos e da própria evolução temporal.

Segundo Marques (2000), a principal vantagem da utilização de modelos de dados em painel refere-se ao controle da heterogeneidade individual, ou seja, à possibilidade de se medirem separadamente os efeitos gerados por conta de diferenças existentes entre cada observação em cada *cross-section*, além de ser possível avaliar a evolução, para um dado indivíduo, das variáveis em estudo ao longo do tempo.

Por outro lado, ainda de acordo com Marques (2000), os dados em painel providenciam maior quantidade de informação, maior variabilidade dos dados, menor colinearidade entre as variáveis, maior número de graus de liberdade e maior eficiência na estimação. A inclusão da dimensão em *cross-section*, num estudo temporal, confere maior variabilidade aos dados, na medida em que a utilização de dados agregados resulta em séries mais suaves do que as séries individuais que lhes servem de base. Este aumento na variabilidade dos dados contribui para a redução de uma eventual colinearidade existente entre variáveis.

Usaremos em nossos exemplos as bases de dados **Painel Curto.dta** e **Painel Longo.dta**. As referidas bases contêm, respectivamente, 11.220 e 580 observações referentes a dados fictícios sobre mortalidade por causas externas ao longo do tempo para municípios provenientes de cinco estados da federação (Quadro 9.1).

Quadro 9.1 Variáveis que compõem as bases de dados **Painel Curto.dta** e **Painel Longo.dta**

Variável	Descrição	Tipo
mês	Mês	
id	Identificação do município	
estado	Estado da federação	Qualitativa
t	Período para cada município	Quantitativa
ano	Ano	
renda	Renda média familiar (R$) do município em determinado mês	Quantitativa
invest	Investimento mensal em segurança pública (R$ x 10.000) no município em determinado mês	Quantitativa
mort	Mortalidade ou causas externas (para cada 100.000 habitantes) no município em determinado mês	Quantitativa

Este capítulo tem como objetivo apresentar e discutir os principais estimadores de dados em painel que podem ser utilizados, bem como auxiliar na definição do modelo mais consistente a ser adotado, em função das características dos dados.

9.1. MODELOS DE DADOS EM PAINEL

Existem muitos modelos diferentes que podem ser utilizados para dados em painel. A distinção básica entre eles, segundo Greene (2007), é a existência de efeitos fixos ou aleatórios. O termo "efeitos fixos" oferece uma ideia equivocada da modelagem uma vez que, para ambos os casos, os efeitos no nível do indivíduo (firmas, entidades governamentais ou países, por exemplo) são aleatórios. Assim, segundo Cameron e Trivedi (2009), os modelos de efeitos fixos apresentam a complicação adicional de que os regressores sejam correlacionados com os efeitos do nível do indivíduo e, portanto, uma estimação consistente dos parâmetros do modelo requer uma eliminação ou controle dos efeitos fixos. Um modelo que leva em conta os efeitos específicos do indivíduo *i* para uma variável dependente y_{it} especifica que:

$$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \varepsilon_{it} \qquad \text{[Equação 9.1]}$$

em que x'_{it} são regressores, β_{0i} são os efeitos aleatórios específicos de indivíduo e ε_{it} representa o erro idiossincrático.

Fazendo o termo do erro ser $\mu_{it} = \beta_{0i} + \varepsilon_{it}$ e permitindo que x'_{it} seja correlacionado com o termo de erro invariante no tempo (β_{0i}), presume-se que x'_{it} não seja correlacionado com o erro idiossincrático ε_{it}. O modelo de efeitos fixos implica que $E(y_{it} | \beta_{0i}, x_{it}) = \beta_{0i} + x'_{it}\beta_1$, presumindo que $E(\varepsilon_{it} | \beta_{0i}, x_{it}) = 0$, de modo que $\beta_j = \partial E(y_{it} | \beta_{0i}, x_{it}) / \partial x_{j,it}$. A vantagem do modelo de efeitos fixos é que pode ser obtido um estimador consistente do efeito marginal do j-ésimo regressor de $E(y_{it} | \beta_{0i}, x_{it})$, dado que $x_{j,it}$ varia no tempo.

No modelo de efeitos aleatórios, por outro lado, pressupõe-se que β_{0i} é puramente aleatório, ou seja, que não é correlacionado com os regressores. A estimação, portanto, é elaborada com um estimador FGLS (*feasible generalized least squares*). A vantagem do modelo de efeitos aleatórios é que este estima todos os coeficientes, mesmo dos regressores invariantes no tempo, e, portanto, os efeitos marginais. Ademais, $E(y_{it} | x_{it})$ pode ser estimado. Porém, a grande desvantagem é que estes estimadores são inconsistentes se o modelo de efeitos fixos for mais apropriado.

Conforme já discutido, a variável dependente e os regressores podem potencialmente variar simultaneamente ao longo do tempo e entre indivíduos. Enquanto a variação, ao longo do tempo ou para um dado indivíduo, é conhecida por *within variance*, a variação entre indivíduos é chamada de *between variance*. De acordo com Wooldridge (2010), no modelo de efeitos fixos o coeficiente de um regressor com baixa variação *within* será imprecisamente estimado e não será identificado se não houver qualquer *within variance*. Assim, é de fundamental importância a distinção entre estas variações para a definição do melhor modelo de dados em painel.

A variação total das observações de um regressor x em torno da média geral $\overline{x} = 1/\sum_i T_i \sum_i \sum_t x_{it}$ no conjunto de dados pode ser decomposta na soma da variação *within* ao longo do tempo para cada indivíduo em torno de $\overline{x}_i = 1/T \sum_t x_{it}$ e na variação *between* entre indivíduos (para $\overline{x}_i$ em torno de $\overline{x}$). De acordo com Cameron e Trivedi (2009):

Variância *Within*: $s^2_{xW} = \frac{1}{\sum_i T_i - 1} \sum_i \sum_t (x_{it} - \overline{x}_i + \overline{x})^2$ [Equação 9.2]

Variância *Between*: $s^2_{xB} = \frac{1}{N-1} \sum_i (\overline{x}_i - \overline{x})^2$ [Equação 9.3]

Variância Geral: $s^2_{xO} = \frac{1}{\sum_i T_i - 1} \sum_i \sum_t (x_{it} - \overline{x})^2$ [Equação 9.4]

As notações N e $\sum_i T_i$ correspondem, respectivamente, ao número de indivíduos e ao número total de observações ao longo do tempo.

Este capítulo traz a aplicação de modelagens com painel de dados por meio de dez diferentes estimadores, a fim de propiciar um melhor entendimento dos seus conceitos e das suas condições de uso. O Quadro 9.2, com base em Cameron e Trivedi (2009) e em Fávero (2013), apresenta estes dez diferentes modelos.

Quadro 9.2 Modelos de dados em painel a serem estimados

Modelo	Descrição
POLS com Erros-Padrão Robustos Clusterizados	$y_{it} = \beta_0 + x'_{it}\beta_1 + \mu_{it}$ Estimação MQO (mínimos quadrados ordinários) com controle da correlação *within* do erro μ_{it} ao longo do tempo.
Modelo com Estimador *Between*	$\overline{y}_i = \beta_0 + x'_i\beta_1 + (\beta_{0i} - \beta_0 + \overline{\varepsilon}_i)$ O estimador *between* somente utiliza a variação das *cross-sections* e é o estimador MQO de uma regressão de $\overline{y}_i$ em função de $\overline{x}_i$. A consistência deste estimador requer que o termo de erro $(\beta_{0i} - \beta_0 + \overline{\varepsilon}_i)$ não seja correlacionado com x_{it}.
Efeitos Fixos	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \varepsilon_{it}$ Os parâmetros β_{0i} podem ser correlacionados com os regressores x_{it}, o que permite uma forma limitada de endogeneidade. Pressupõe-se que x_{it} não seja correlacionado com o erro idiossincrático ε_{it}.
Efeitos Fixos com Erros-Padrão Robustos Clusterizados	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \varepsilon_{it}$ Os termos β_{0i} podem ser correlacionados com os regressores x_{it}, o que permite uma forma limitada de endogeneidade. Presume-se que os erros sejam independentes entre indivíduos e que ε_{it} seja heterocedástico.

Quadro 9.2 Modelos de dados em painel a serem estimados *(cont.)*

Modelo	Descrição
Efeitos Aleatórios	$y_{it} = x'_{it}\beta_1 + (\beta_{0i} + \varepsilon_{it})$ Os parâmetros β_{0i} e os termos de erro idiossincrático ε_{it} são independentes e identicamente distribuídos (i.i.d.). O estimador de efeitos aleatórios é o FGLS de β_1, dado que $corr(\mu_{it}, \mu_{is}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$.
Efeitos Aleatórios com Erros-Padrão Robustos Clusterizados	$y_{it} = x'_{it}\beta_1 + (\beta_{0i} + \varepsilon_{it})$ Se não houver efeitos fixos, mas os erros apresentarem correlação *within*, o estimador de efeitos aleatórios é consistente, porém ineficiente. Portanto, erros-padrão robustos clusterizados precisam ser obtidos.
Efeitos Fixos com Termos de Erro AR(1)	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \mu_{it}$ Com $\mu_{it} = \rho_i \mu_{i,t-1} + \varepsilon_{it}$. Considera-se β_{0i} como sendo um efeito fixo.
Efeitos Aleatórios com Termos de Erro AR(1)	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \mu_{it}$ Com $y_{it} = \rho_i \mu_{i,t-1} + \varepsilon_{it}$. Considera-se β_{0i} como sendo um efeito aleatório.
Pooled com Método de Estimação MQO e Termos de Erro AR(1)	$y_{it} = \beta_{0i} + \gamma_t + x'_{it}\beta_1 + \varepsilon_{it}$ Com $\mu_{it} = \rho_i \mu_{i,t-1} + \varepsilon_{it}$, em que os ε_{it} são serialmente não correlacionados, mas com correlação entre indivíduos igual a $corr(\varepsilon_{it}, \varepsilon_{is}) = \sigma_{ts} \neq 0$.
Pooled com Método de Estimação FGLS e Termos de Erro AR(1)	$y_{it} = \beta_{0i} + \gamma_t + x'_{it}\beta_1 + \varepsilon_{it}$ Similar ao modelo *pooled* com método de estimação MQO, mas com estimador FGLS.

9.2. APLICAÇÃO

Como muitas bases de dados em ciências sociais aplicadas apresentam periodicidade de divulgação mensal, trimestral ou anual, é comum que os estudos nestas áreas utilizem modelos de dados em painel curto, já que o número de indivíduos (empresas, municípios ou países, por exemplo) ultrapassa o número de períodos de divulgação dos dados. Por outro lado, nada impede que o pesquisador baseie seu estudo numa amostra menor de indivíduos ou utilize dados com frequência de divulgação maior (diária, por exemplo) o que poderia ocasionar uma modelagem com dados em painel longo. De qualquer maneira, é fundamental que a identificação desta característica na base de dados seja feita de forma anterior à modelagem propriamente dita.

Inicialmente, uma base fictícia contendo dados sobre mortalidade por causas externas para cada 100.000 habitantes (indicador de violência) em 299 municípios provenientes de 5 estados brasileiros (Bahia, Goiás, Minas Gerais, Pará e São Paulo), ao longo de 100 meses (2006-2012), totalizando 11.220 observações, será utilizada para o estudo de um painel curto (arquivo **Painel Curto.dta**). Na sequência, um estrato desta base será utilizado, com dados de apenas 10 municípios ao longo de 58 meses,

totalizando 580 observações, com o objetivo de se estudar o painel longo (arquivo **Painel Longo.dta**).

A definição dos indivíduos (municípios) e dos períodos (meses) é dada pelo comando:

xtset id t

RESULTADOS 9.1 Definindo o painel.

```
. xtset id t
        panel variable:  id (unbalanced)
         time variable:  t, 1 to 100, but with gaps
                 delta:  1 unit
```

A base apresenta dados considerados desbalanceados, uma vez que não há uma quantidade igual de períodos para cada um dos municípios estudados.

Antes de elaborarmos os modelos de regressão em painel propriamente ditos, iremos analisar o comportamento da mortalidade por causas externas ao longo do tempo. Por meio da Figura 9.1, é possível verificar que este indicador de violência urbana apresenta compor-

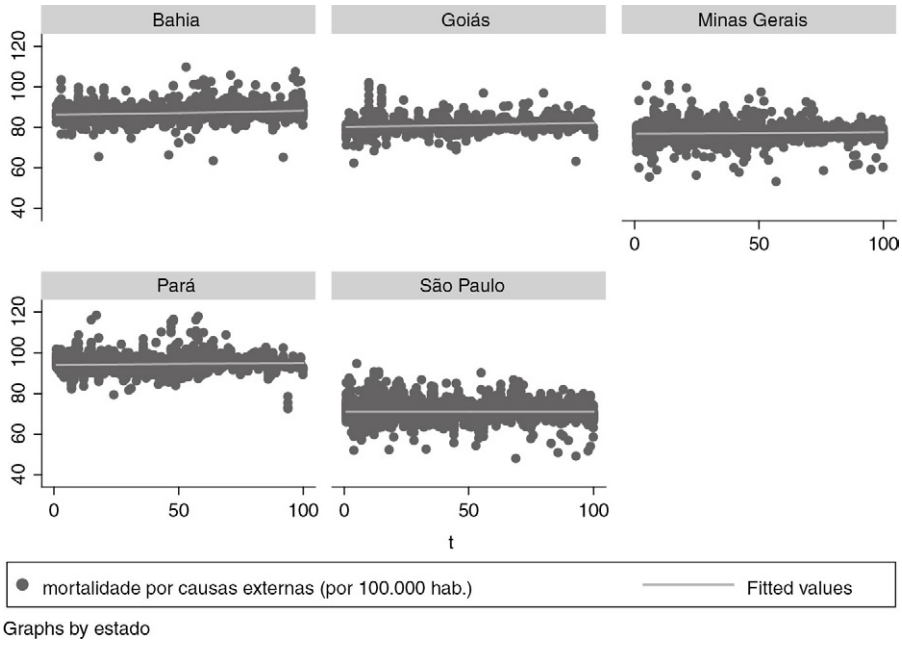


Figura 9.1 ***Evolução da mortalidade por causas externas para os municípios de cada estado.***

tamento distinto, em média, para cada um dos 5 estados brasileiros ao longo do tempo. Apesar de a análise ser feita para cada município, a Figura 9.1, obtida por meio do comando a seguir, apresenta o comportamento para todos os municípios de cada estado.

graph twoway scatter mort t || lfit mort t, by(estado)

RESULTADOS 9.2 Gerando o gráfico de mortalidade em função do tempo para cada estado.

```
. graph twoway scatter mort t || lfit mort t, by(estado)
```

Cada ponto na Figura 9.1 representa um par mortalidade-mês para determinado município. Enquanto alguns estados apresentam crescimentos neste indicador de violência, outros apresentam, ainda que de forma incipiente, certa redução. Este comportamento sugere a elaboração de modelos longitudinais, já que as razões que levam a este fenômeno (regressores) podem variar entre municípios e ao longo do tempo, conforme será apresentado e discutido adiante. Enquanto a Figura 9.2 apresenta a variação dos indicadores de mortalidade por causas externas ao longo do tempo para cada município, ou seja, mostra os desvios do indicador de violência em relação à média individual de cada município (*within variation*), a Figura 9.3 apresenta a

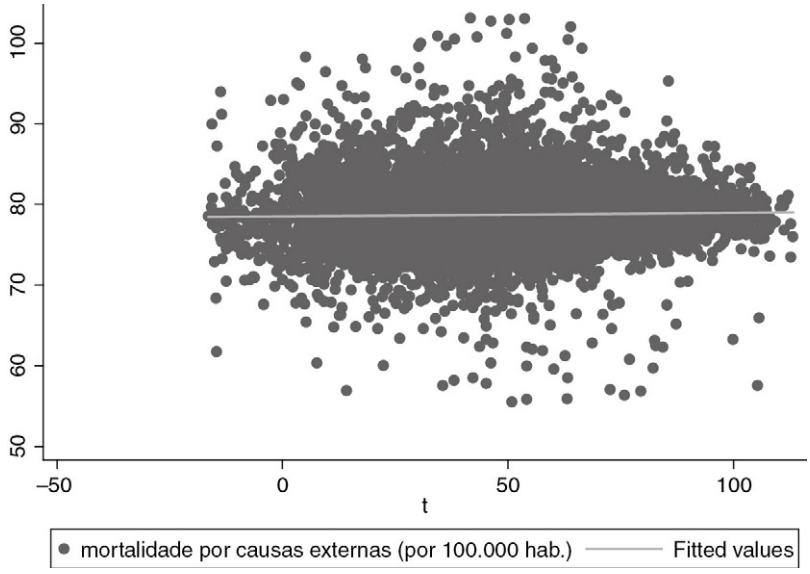


Figura 9.2 ***Desvios da mortalidade por causas externas em relação à média de cada município ao longo do tempo (*within variation*).***

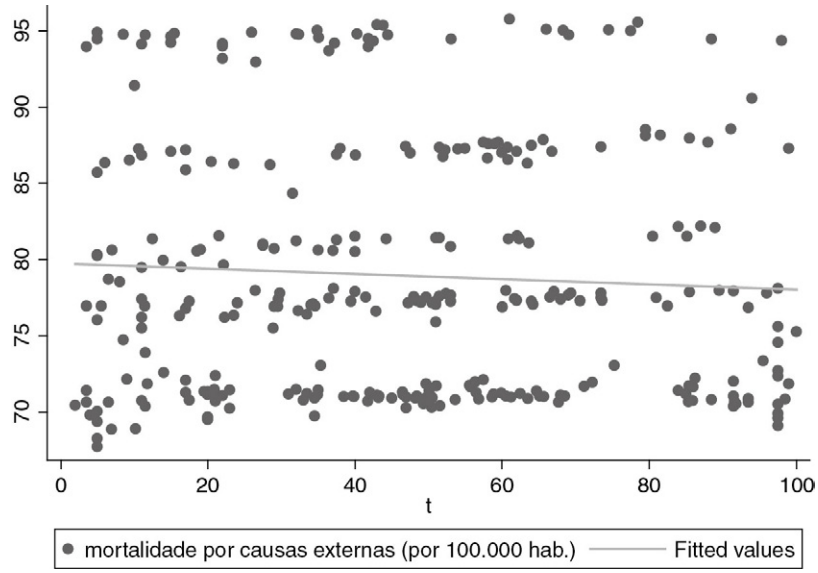


Figura 9.3 ***Desvios da mortalidade por causas externas em relação à média geral para cada instante de tempo (*between variation*).***

variação deste indicador entre os municípios, ou seja, mostra os desvios da mortalidade por causas externas nos municípios em relação à média geral para cada instante de tempo (*between variation*). Os comandos para a elaboração das Figuras 9.2 e 9.3 são, respectivamente:

```
preserve
xtdata, fe
graph twoway scatter mort t || lfit mort t
restore
```

RESULTADOS 9.3 Gerando o gráfico de *within variation*.

```
. preserve
. xtdata, fe
. graph twoway scatter mort t || lfit mort t
. restore
```

```
preserve
xtdata, be
graph twoway scatter mort t || lfit mort t
restore
```

RESULTADOS 9.4 Gerando o gráfico de *between variation*.

```
. preserve
. xtdata, be
. graph twoway scatter mort t || lfit mort t
. restore
```

Por meio da Figura 9.3 é possível verificar que há 5 patamares de mortalidade por causas externas, correspondentes aos municípios situados em cada um dos 5 estados considerados na análise. Os Resultados 9.5, obtidos por meio do comando **tabstat**, possibilitam que observemos este fenômeno.

tabstat mort, by(estado)

Como discutido, 10 diferentes modelos de dados em painel serão elaborados, com diferentes considerações sobre os estimadores e os termos de erro. O modelo geral é dado por:

$$mort_{it} = \beta_{0i} + \beta_1.(renda)_{it} + \beta_2.(invest)_{it} + \varepsilon_{it} \qquad \text{[Equação 9.5]}$$

em que β_1 e β_2 representam as mudanças na mortalidade mensal por causas externas para cada 100.000 habitantes quando uma unidade de renda média familiar mensal (R$) é gerada, ou quando uma unidade de investimento mensal em segurança pública (R$ x 10.000) é disponibilizada, respectivamente, mantidas as demais condições constantes.

A seguir, serão discutidos os resultados das modelagens, tanto para um painel curto, quanto para um painel longo.

RESULTADOS 9.5 Obtendo as médias da variável *mort* por *estado*.

```
. tabstat mort, by(estado)

Summary for variables: mort
     by categories of: estado

        estado |      mean
---------------+----------
         Bahia |   87.2779
         Goiás |  81.05629
  Minas Gerais |  77.24532
          Pará |  94.49694
     São Paulo |  71.10612
---------------+----------
         Total |  78.71848
-------------------------
```

9.2.1 Modelos para dados em painel curto

Como a amostra, neste caso, oferece dados de 299 municípios em 100 meses, o painel pode ser considerado curto (T < N).

A seguir é apresentada a decomposição de variância para cada uma das variáveis, obtida por meio do comando **xtsum**.

xtsum id t mort renda invest

RESULTADOS 9.6 Obtendo a decomposição de variância para cada variável.

```
. xtsum id t mort renda invest

Variable         |       Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
id       overall |  246204.2   29044.95      213100     293290 |     N =   11220
         between |             29226.07      213100     293290 |     n =     299
         within  |                   0    246204.2   246204.2 |  T-bar = 37.5251
                 |                                            |
t        overall |  47.70481   29.18401           1        100 |     N =   11220
         between |             27.49955           2        100 |     n =     299
         within  |             21.47196   -16.42019   113.0894 |  T-bar = 37.5251
                 |                                            |
mort     overall |  78.71848   8.739225    48.01177   118.4197 |     N =   11220
         between |             8.143473    67.75149   95.76708 |     n =     299
         within  |             3.713448    55.50703   103.0668 |  T-bar = 37.5251
                 |                                            |
renda    overall |   3708.86   704.7227      270.77   5489.623 |     N =   11220
         between |             720.6189    2256.562   4378.631 |     n =     299
         within  |             49.38064    1618.276   4819.852 |  T-bar = 37.5251
                 |                                            |
invest   overall |  515.4366    116.598    30.96548   691.8675 |     N =   11220
         between |             118.8921    311.1364   640.0215 |     n =     299
         within  |             7.395163    73.60609   696.2087 |  T-bar = 37.5251
```

De acordo com os Resultados 9.6, nota-se que o município é obviamente invariante ao longo do tempo e, portanto, apresenta variação *within* igual a zero. Por outro lado, a variável referente ao tempo *t* não é invariante entre municípios, já que se trata de um painel desbalanceado e, portanto, a sua variação *between* não é igual a zero. Todas as variáveis da Equação 9.5 apresentam maior variação entre municípios (*between*) do que ao longo do tempo (*within*), porém ainda não é possível afirmar que a estimação *within* resultará numa perda de eficiência, já que a proporção entre as variâncias *within* e *between* de cada variável é diferente e ainda não se conhecem as significâncias estatísticas de cada uma delas nos modelos. Os resultados obtidos por meio do comando **xtsum**, todavia, oferecem maior embasamento para a adoção dos modelos de dados em painel e a aplicação de diversos estimadores. As colunas "Mínimo" e "Máximo" apresentam, respectivamente, os valores mínimos e máximos de x_{it} para a linha "geral", $\overline{x}_i$ para a linha "*between*" e $\left(x_{it} - \overline{x}_i + \overline{x}\right)$ para a linha "*within*".

Dessa forma, partiremos agora para a elaboração das diversas regressões para o painel curto. Os comandos para a realização de cada uma delas encontram-se a seguir:

- POLS com Erros-Padrão Robustos Clusterizados:

reg mort renda invest, vce(cluster id)

- Modelo com Estimador *Between*:

xtreg mort renda invest, be

- Efeitos Fixos:

xtreg mort renda invest, fe

- Efeitos Fixos com Erros-Padrão Robustos Clusterizados:

xtreg mort renda invest, fe vce(cluster id)

- Efeitos Aleatórios:

xtreg mort renda invest, re

- Efeitos Aleatórios com Erros-Padrão Robustos Clusterizados:

xtreg mort renda invest, re vce(cluster id)

Os Resultados 9.7 apresentam os *outputs* dos seis modelos de dados em painel curto, gerados por meio do seguinte comando:

```
quietly reg mort renda invest, vce(cluster id)
estimates store POLS_rob
quietly xtreg mort renda invest, be
estimates store BE
quietly xtreg mort renda invest, fe
estimates store FE
quietly xtreg mort renda invest, fe vce(cluster id)
estimates store FE_rob
quietly xtreg mort renda invest, re
estimates store RE
quietly xtreg mort renda invest, re vce(cluster id)
estimates store RE_rob
estimates table POLS_rob BE FE FE_rob RE RE_rob, b se stats(N r2 r2_o r2_b r2_w F chi2)
```

Como se pode observar, os coeficientes estimados variam de modelo para modelo, o que reflete a existência de resultados diferentes se as variações *within* ou *between* forem utilizadas.

Primeiramente verifica-se, em relação à adequação dos modelos, que o vetor de regressores apresenta significância estatística em todos os casos (sig. F para os modelos POLS, *between* e com efeitos fixos, e sig. Wald χ^2 para os modelos com efeitos aleatórios). Além disso, verifica-se a existência de maiores valores para os R^2 *between* em todos os modelos em que esta estatística é calculada, o que demonstra que a variação que ocorre na variável dependente é consideravelmente maior entre os municípios do que para um mesmo

RESULTADOS 9.7 Apresentando os *outputs* dos modelos em painel curto.

```
. quietly reg mort renda invest, vce(cluster id)
. estimates store POLS_rob

. quietly xtreg mort renda invest, be
. estimates store BE

. quietly xtreg mort renda invest, fe
. estimates store FE

. quietly xtreg mort renda invest, fe vce(cluster id)
. estimates store FE_rob

. quietly xtreg mort renda invest, re
. estimates store RE

. quietly xtreg mort renda invest, re vce(cluster id)
. estimates store RE_rob

. estimates table POLS_rob BE FE FE_rob RE RE_rob, b se stats(N r2 r2_o r2_b r2_w F chi2)

-------------------------------------------------------------------------------------
    Variable |  POLS_rob         BE          FE        FE_rob         RE         RE_rob
-------------+-----------------------------------------------------------------------
       renda | -.00650285   -.00727341   -.00136705   -.00136705   -.00542681   -.00542681
             |  .00071806    .00054628     .0007951    .00062781    .00034097    .00099905
      invest | -.02835148   -.02349384   -.02268309   -.02268309   -.03471683   -.03471683
             |  .00432726    .00331108    .00530924    .00465684    .00207878    .00594336
       _cons |  117.45004    117.68013     95.48036     95.48036    116.65975    116.65975
             |  .53752686    .55076075    4.8035626    4.2436396     .4597037    .74389982
-------------+-----------------------------------------------------------------------
           N |       11220         11220        11220        11220        11220        11220
          r2 |    .8018893     .96071666    .00167011    .00167011
        r2_o |                  .8016247    .79669279    .79669279    .80141038    .80141038
        r2_b |                 .96071666     .9524374     .9524374    .95917397    .95917397
        r2_w |                 .00035193    .00167011    .00167011    .00109736    .00109736
           F |  8370.7721     3619.5006     9.133222    12.368975
        chi2 |                                                      8424.851     12847.511
-------------------------------------------------------------------------------------
                                                                              legend: b/se
```

município ao longo do tempo. Em outras palavras, a mortalidade por causas externas não tem se alterado em média ao longo do tempo para cada um dos municípios estudados. Entretanto, seus valores médios são diferentes quando a comparação é elaborada entre os municípios.

Com relação aos regressores (variáveis *renda* e *invest*), verifica-se, para todos os modelos, que os respectivos coeficientes são estatisticamente diferentes de zero. O mesmo também pode ser dito em relação ao intercepto.

Os regressores estimados para o modelo de efeitos aleatórios oferecem erros-padrão, que são apresentados abaixo do coeficiente de cada regressor para cada modelo, menores do que para qualquer outro modelo. O teste Breusch-Pagan, cujo comando é aplicado após a modelagem de efeitos aleatórios (comando **xttest0**), auxilia na rejeição da hipótese nula de que há adequação do modelo POLS em relação ao modelo de efeitos aleatórios, já que $\chi^2 = 741{,}84$ (sig. $\chi^2 = 0{,}000$).

xttest0

RESULTADOS 9.8 Teste Breusch-Pagan.

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        mort[id,t] = Xb + u[id] + e[id,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                 ---------+-----------------------------
                    mort |   76.37405        8.739225
                       e |   14.14491        3.760971
                       u |   1.649652        1.284388

        Test:   Var(u) = 0
                                 chibar2(01) =   741.84
                              Prob > chibar2 =   0.0000
```

Na sequência, por meio do teste F de Chow, que é apresentado ao se estimar o modelo de efeitos fixos, rejeita-se a hipótese nula de que há igualdade de interceptos e inclinações para todos os municípios (POLS). Portanto, estes parâmetros diferem daqueles obtidos por meio do modelo de efeitos fixos, já que F = 3,63 (sig. F = 0,000).

xtreg mort renda invest, fe

RESULTADOS 9.9 Modelo de efeitos fixos, com destaque para o teste F de Chow.

```
. xtreg mort renda invest, fe

Fixed-effects (within) regression               Number of obs      =     11220
Group variable: id                              Number of groups   =       299

R-sq:  within  = 0.0017                         Obs per group: min =         1
       between = 0.9524                                        avg =      37.5
       overall = 0.7967                                        max =       100

                                                F(2,10919)         =      9.13
corr(u_i, Xb)  = 0.9539                         Prob > F           =    0.0001

------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       renda |   -.001367   .0007951    -1.72   0.086    -.0029256    .0001915
      invest |  -.0226831   .0053092    -4.27   0.000    -.0330902    -.012276
       _cons |   95.48036   4.803563    19.88   0.000     86.06451    104.8962
-------------+----------------------------------------------------------------
     sigma_u |   4.639705
     sigma_e |  3.7609715
         rho |  .60347056   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(298, 10919) =     3.63          Prob > F = 0.0000
```

Por fim, segundo Cameron e Trivedi (2009), é essencial que se discuta a distinção entre os modelos de efeitos fixos e aleatórios na análise de dados em painel. Sob a hipótese nula de que os efeitos individuais são aleatórios, o teste de Hausman verifica se os estimadores são similares (efeitos aleatórios) ou divergem entre si (efeitos fixos) para cada indivíduo. Já discutimos, quando da aplicação do comando **xtsum** e por meio da análise do R^2 *within* e do R^2 *between*, que pouca variação ocorre na variável dependente ao longo do tempo para cada município (R^2 *within* baixo e bem menor do que o R^2 *between*), porém alterações visíveis são percebidas entre indivíduos. Neste momento, portanto, é importante saber se os estimadores que influenciam o comportamento da variável dependente entre municípios também divergem entre municípios (efeitos fixos).

No nosso exemplo, a aplicação do teste de Hausman (comando apresentado a seguir) auxilia na rejeição da hipótese nula de que o modelo de efeitos aleatórios oferece estimativas dos parâmetros mais consistentes, já que, para este caso, $\chi^2 = 36{,}53$ (sig. $\chi^2 = 0{,}000$), conforme mostram os Resultados 9.10.

hausman FE RE, sigmamore

RESULTADOS 9.10 Teste de Hausman.

```
. hausman FE RE, sigmamore

                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
      renda  |   -.001367    -.0054268        .0040598        .0007176
     invest  |  -.0226831    -.0347168        .0120337        .0048808
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtreg
           B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                 chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                         =       36.53
               Prob>chi2 =      0.0000
```

Desta forma, seguimos com o modelo de efeitos fixos.

É interessante comentar que, como não há alterações significativas na mortalidade por causas externas para cada município ao longo do tempo, se as regressões tivessem sido elaboradas apenas com o tempo (variável *t*) como regressor da variável *mort*, o teste de Hausman não rejeitaria a hipótese nula de que os efeitos individuais fossem aleatórios, ou seja, o estimador da variável *t* seria similar entre todos os indivíduos.

Seguindo uma importante discussão elaborada por Islam (1995), a principal utilidade da modelagem de dados em painel é permitir que sejam analisadas as diferenças que porventura ocorram entre empresas, setores, municípios, estados, países, entre outras

classificações. Os Resultados 9.11 apresentam os coeficientes da regressão de dados em painel com efeitos fixos para cada um dos estados da amostra.

```
preserve
statsby, by(estado) clear: xtreg mort renda invest, fe
list, clean
restore
```

Embora o indicador de violência urbana (mortalidade por causas externas) sofra influência negativa da evolução da renda média familiar mensal e do montante mensal disponibilizado para investimento em segurança pública nos municípios, verifica-se que essas influências ocorrem de forma diferente e, em algumas localidades, inclusive com sinal invertido em relação à média geral. Os diferentes coeficientes e sinais dos regressores e da constante expressam a importância de se considerar a modelagem de dados em painel e propiciam a formulação de novos estudos.

RESULTADOS 9.11 Coeficientes da regressão em painel com efeitos fixos para cada estado.

```
. preserve

. statsby, by(estado) clear: xtreg mort renda invest, fe
(running xtreg on estimation sample)

      command:  xtreg mort renda invest, fe
           by:  estado

Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....

. list, clean

             estado     _b_renda    _b_invest     _b_cons
  1.          Bahia    -.023373     .0594407    133.6311
  2.          Goiás   -.0091834    -.1119817    159.4829
  3.   Minas Gerais    -.002943    -.0264067     102.875
  4.           Pará    .0028669    -.0568054    105.6951
  5.      São Paulo    .0007185     .0188598    55.94058

. restore
```

9.2.2 Modelos para dados em painel longo

Para este caso, como a amostra oferece dados de 10 municípios ao longo de 58 meses para cada um deles (painel balanceado), o painel pode ser considerado longo (T > N). Aplicaremos o mesmo comando **xtset** para que seja definido o painel.

xtset id t

RESULTADOS 9.12 Definindo o painel longo.

```
. xtset id t
       panel variable:  id (strongly balanced)
        time variable:  t, 1 to 58
                delta:  1 unit
```

Como a influência temporal é bastante importante em séries longas, modelos de efeitos fixos e aleatórios serão também aplicados com a consideração de componentes auto-regressivos (AR(1)) para os resíduos, o que pode resultar em parâmetros com estimativas mais eficientes para painéis longos.

Assim como elaborado para o painel curto, os Resultados 9.13 apresentam a decomposição de variância para cada uma das variáveis do painel longo.

xtsum id t renda mort invest

RESULTADOS 9.13 Obtendo a decomposição de variância para cada variável.

```
. xtsum id t renda mort invest

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
id       overall |  247181.8   30742.95      213100     293168 |     N =     580
         between |             32377.97      213100     293168 |     n =      10
         within  |                   0    247181.8   247181.8 |     T =      58
                 |                                            |
t        overall |      29.5   16.75512           1         58 |     N =     580
         between |                   0        29.5       29.5 |     n =      10
         within  |             16.75512           1         58 |     T =      58
                 |                                            |
renda    overall |  3381.312   727.7594    2232.427   4349.093 |     N =     580
         between |              766.443    2274.937   4333.674 |     n =      10
         within  |             5.399736    3338.803   3396.732 |     T =      58
                 |                                            |
mort     overall |  77.90436   8.791768    53.99449    111.661 |     N =     580
         between |             8.559425    66.21396   90.79237 |     n =      10
         within  |             3.353194    62.46154   98.77302 |     T =      58
                 |                                            |
invest   overall |  459.5017   113.3427     311.799   639.6788 |     N =     580
         between |             119.3706    312.0222   639.0717 |     n =      10
         within  |             .1166738    458.8273   460.1538 |     T =      58
```

Verifica-se que as variáveis *mort*, *renda* e *invest* apresentaram maior variação entre indivíduos (*between*) do que ao longo do tempo (*within*). Por outro lado, a variável temporal (*t*) passa a apresentar variação *between* nula, já que se trata de um painel balanceado.

Da mesma forma que o procedimento realizado para o painel curto, os Resultados 9.14 apresentam os *outputs* dos modelos, considerando também seis diferentes estimadores. Os comandos para a realização de cada um deles isoladamente são:

- Efeitos Fixos:

xtreg mort renda invest, fe

- Efeitos Aleatórios:

xtreg mort renda invest, re

- Efeitos Fixos com Erros AR(1):

xtregar mort renda invest, fe

- Efeitos Aleatórios com Erros AR(1):

xtregar mort renda invest, re

- POLS com Erros AR(1) e correlação entre indivíduos:

xtpcse mort renda invest, corr(ar1)

- FGLS com Erros AR(1) e correlação entre indivíduos:

xtgls mort renda invest, corr(ar1) panels(correlated)

Os Resultados 9.14 já apresentam os *outputs* consolidados, obtidos por meio do seguinte comando:

```
quietly xtreg mort renda invest, fe
estimates store FE
quietly xtreg mort renda invest, re
estimates store RE
quietly xtregar mort renda invest, fe
estimates store FEAR1
quietly xtregar mort renda invest, re
estimates store REAR1
quietly xtpcse mort renda invest, corr(ar1)
estimates store POLSAR1
quietly xtgls mort renda invest, corr(ar1) panels(correlated)
estimates store FGLSAR1
estimates table FE RE FEAR1 REAR1 POLSAR1 FGLSAR1, b se stats(N r2 r2_o r2_b r2_w F chi2)
```

De acordo com os Resultados 9.14, é possível verificar que os coeficientes estimados também variam entre os modelos. Ao se permitir que os termos de erro sejam correlacionados entre municípios, verifica-se que ocorre, nesse exemplo, uma redução dos erros-padrão dos modelos *pooled* com estimadores MQO e FGLS em comparação com aqueles obtidos anteriormente por meio dos modelos de efeitos fixos e aleatórios com termos de erro AR(1).

Em relação à adequação dos modelos propriamente ditos, nota-se a significância estatística do conjunto de variáveis em todos os casos, à exceção dos modelos com efeitos fixos com e sem termos de erro AR(1). Todavia, apenas nos modelos *pooled* com estimadores MQO e FGLS os regressores *renda* e *invest* são estatisticamente significantes, a um

RESULTADOS 9.14 Apresentando os *outputs* dos modelos em painel longo.

```
. quietly xtreg mort renda invest, fe
. estimates store FE

. quietly xtreg mort renda invest, re
. estimates store RE

. quietly xtregar mort renda invest, fe
. estimates store FEAR1

. quietly xtregar mort renda invest, re
. estimates store REAR1

. quietly xtpcse mort renda invest, corr(ar1)
. estimates store POLSAR1

. quietly xtgls mort renda invest, corr(ar1) panels(correlated)
. estimates store FGLSAR1

. estimates table FE RE FEAR1 REAR1 POLSAR1 FGLSAR1, b se stats(N r2 r2_o r2_b r2_w F chi2)

-------------------------------------------------------------------------------------
    Variable |      FE           RE          FEAR1         REAR1        POLSAR1      FGLSAR1
-------------+-----------------------------------------------------------------------
       renda |   .03208082   -.00868814     .03523597   -.00881594    -.00905159   -.00841292
             |   .02675909    .00268194     .02697191    .00225389     .0010679    .0009141
      invest |   .64134944   -.01545138     .78883148   -.01464767    -.01315997   -.01640914
             |   1.2384276    .01722392     1.2552555    .01447386     .00660736    .0056251
       _cons |  -325.27208     114.3816    -403.69665    114.44421     114.55634    114.01692
             |   596.67366    2.1759433     595.8692    1.8265573     .90538354    .72671864
-------------+-----------------------------------------------------------------------
           N |         580           580           570           580           580          580
          r2 |   .00262936                                                .81978445
        r2_o |   .82669735    .83890377     .82839386    .83892521
        r2_b |   .96750407    .98196673     .96820013    .98199588
        r2_w |   .00262936    .00223196     .00325154    .00222724
           F |   .74870614                  .9101378
        chi2 |                381.2044                   542.07109     2835.4893    4269.6126
-------------------------------------------------------------------------------------
                                                                                legend: b/se
```

nível de 5% de significância, para explicar o comportamento da variável dependente. Para este último modelo (*pooled* com método de estimação FGLS e termos de erro AR(1)), os parâmetros dos regressores são ainda mais significantes, uma vez que os erros-padrão são consideravelmente mais baixos.

Para dados em painel longo, a consideração de efeitos individuais com termos de erro AR(1) pode resultar em modelos melhores do que se forem considerados termos de erro i.i.d., o que poderá gerar estimativas dos parâmetros mais eficientes, como ocorre neste caso.

9.3. CONSIDERAÇÕES FINAIS

Modelos de dados em painel possibilitam que o pesquisador avalie a relação entre alguma variável de desempenho e diversas variáveis preditoras, permitindo que se elaborem inferências sobre as eventuais diferenças entre indivíduos e ao longo do tempo a respeito da evolução daquilo que se pretende estudar. Dadas as suas características, é natural que muitas pesquisas em ciências sociais aplicadas venham a fazer uso de tais modelos, uma vez que muitos dados são publicados com determinada periodicidade para empresas, municípios, estados ou países.

Para tanto, é necessário, assim como para qualquer outra técnica de modelagem, que a aplicação venha acompanhada de rigor metodológico e certos cuidados quando da análise dos resultados, principalmente se estes tiverem como objetivo a elaboração de previsões. A adoção de determinado estimador, em detrimento de outro considerado viesado ou inconsistente, pode auxiliar o pesquisador na escolha do melhor modelo, valorizando a sua pesquisa e propiciando novos estudos sobre o tema escolhido.

Neste capítulo, procurou-se elaborar seis diferentes modelos para um específico painel curto e outros seis para um painel longo. A análise da contribuição da renda média familiar e do investimento em segurança pública sobre a mortalidade por causas externas de municípios brasileiros possibilita que seja incrementada a discussão sobre violência urbana e desenvolvimento social, porém foi adotada apenas como exemplo dentro de um objetivo específico, que foi o de apresentar como os diferentes estimadores podem gerar resultados discrepantes quando da elaboração de modelos de dados em painel e auxiliar para a escolha do modelo mais adequado, tanto no caso de um painel curto, quanto no de um painel longo.

9.4. EXERCÍCIO

1. Um cardiologista tem monitorado 10 pacientes, que são executivos de empresas, ao longo dos últimos 5 anos, em relação aos seus níveis de colesterol LDL (mg/dL). Seu intuito é orientá-los sobre a importância da manutenção ou perda de peso e da realização periódica de atividades físicas para a redução do colesterol e, portanto, elaborou uma base de dados que pode ser acessada por meio do arquivo **colest.dta**. As variáveis presentes nesta base são:

Variável	Descrição
ano	Ano
indivíduo	Identificação do executivo
colesterol	Colesterol LDL (mg/dL)
imc	Índice de massa corpórea (kg/m^2)
esporte	Atividades físicas semanais (número de vezes)

Por meio do uso desta base de dados, pede-se:

a. Defina o painel com as variáveis *indivíduo* e *ano*. Trata-se de um painel balanceado?

b. Elabore um gráfico que apresenta a evolução do índice de colesterol LDL ao longo dos anos, discriminando cada um dos executivos. É possível, ainda que visualmente, perceber se há diferenças entre o comportamento da evolução anual do índice de colesterol LDL entre os indivíduos?

c. Elabore a decomposição de variância para cada variável e discuta os resultados em termos de variação *within* e *between*.

d. Deseja-se desenvolver o seguinte modelo, a fim de que seja possível verificar a importância da evolução do índice de massa corpórea e da realização de atividades físicas periódicas sobre o índice de colesterol LDL.

$$colesterol_{it} = \beta_{0i} + \beta_1.(imc)_{it} + \beta_2.(esporte)_{it} + \varepsilon_{it}$$

Desta forma, elabore as seguintes estimações, por meio do painel de dados, e discuta os resultados:

- POLS com Erros-Padrão Robustos Clusterizados.
- Modelo com Estimador *Between*.
- Efeitos Fixos.
- Efeitos Fixos com Erros-Padrão Robustos Clusterizados.
- Efeitos Aleatórios.
- Efeitos Aleatórios com Erros-Padrão Robustos Clusterizados.

e. É possível verificar, em relação à adequação dos modelos, que o vetor de regressores apresenta significância estatística em todos os casos (sig. F para os modelos POLS, *between* e com efeitos fixos, e sig. Wald χ^2 para os modelos com efeitos aleatórios)?

f. Verifica-se que os valores de R^2 *between* são maiores do que os valores de R^2 *within* em todos os modelos em que estas estatísticas são calculadas. Justifique por qual razão este fato deve ter ocorrido.

g. Elabore o teste Breusch-Pagan, o teste F de Chow e o teste de Hausman e discuta seus resultados. O que se pode avaliar sobre os efeitos fixos e os efeitos aleatórios neste painel de dados?

h. Elabore uma tabela com os coeficientes do modelo com efeitos fixos para cada um dos executivos da amostra. Há diferenças entre eles, em termos de comportamento das variáveis *imc* e *esporte* sobre a variável *colesterol*? Como você, cardiologista, orientaria cada um dos pacientes?

REFERÊNCIAS

ACOCK, A. C. *A Gentle Introduction to Stata*. 2. ed. College Station: StataCorp LP, 2008.

AHN, S. C.; SCHMIDT, P. Efficient estimation of dynamic panel data models: alternative assumptions and simplified estimation. *Journal of Econometrics*, v. 76, n. 1-2, p. 309-321, 1997.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. *Estatística Aplicada à Administração e Economia*. São Paulo: Pioneira Thomson Learning, 2002.

ANDERSON, T. W.; HSIAO, C. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, v. 18, n. 1, p. 47-82, 1982.

ARELLANO, M. *Panel Data Econometrics*: Advanced Texts in Econometrics. New York: Oxford University Press, 2003.

BALESTRA, P.; NERLOVE, M. Pooling cross section and time series data in the estimation of a dynamic model: the demand for natural gas. *Econometrica*, v. 34, n. 3, p. 585-612, 1966.

BALTAGI, B. H. *Econometric Analysis of Panel Data*. 4. ed. New York: John Wiley & Sons, 2008.

BALTAGI, B. H.; GRIFFIN, J. M. Short and long run effects in pooled models. *International Economic Review*, v. 25, n. 3, p. 631-645, 1984.

BARNETT, V.; LEWIS, T. *Outliers in Statistical Data*. 2. ed. New York: John Wiley & Sons, 1984.

BAUM, C. F. *An Introduction to Modern Econometrics Using Stata*. College Station, Tex: Stata Press, 2006.

BECK, N.; KATZ, J. N. What to do (and not to do) with time-series cross-section data. *American Political Science Review*, v. 89, n. 3, p. 634-647, 1995.

BELKAOUI, A. *Quantitative Models in Accounting*. Quorum Books, 1987.

BERENSON, M. L.; LEVINE, D. M. *Basic Business Statistics*: Concepts and Application. 6. ed. Upper Saddle River: Prentice Hall, 1996.

BHARGAVA, A.; FRANZINI, L.; NARENDRANATHAN, W. Serial correlation and the fixed effects model. *Review of Economic Studies*, v. 49, n. 4, p. 533-549, 1982.

BHARGAVA, A.; SARGAN, J. D. Estimating dynamic random effects models from panel data covering short time periods. *Econometrica*, v. 51, n. 6, p. 1635-1659, 1983.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis:* Forecasting and Control. 4. ed. Hoboken: John Wiley & Sons, 2008.

BREUSCH, T. S.; MIZON, G. E.; SCHMIDT, P. Efficient estimation using panel data. *Econometrica*, v. 57, n. 3, p. 695-700, 1989.

BUENO, R. L. S. *Econometria de Séries Temporais*. 2. ed. São Paulo: Cengage Learning, 2011.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 7. ed. São Paulo: Saraiva, 2011.

CAMERON, A. C.; TRIVEDI, P. K. *Microeconometrics Using Stata*. College Station: Stata Press, 2009.

CHARNET, R.; BONVINO, H.; FREIRE, C. A. L.; CHARNET, E. M. R. *Análise de Modelos de Regressão Linear:* Com Aplicações. 2. ed. Campinas: Editora da UNICAMP, 2008.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, v. 31, p. 93-96, 1951.

DILLON, W. R.; GOLDSTEIN, M. *Multivariate Analysis Methods and Applications*. New York: John Wiley & Sons, 1984.

DOANE, D. P.; SEWARD, L. E. *Estatística Aplicada à Administração e à Economia*. São Paulo: McGraw-Hill, 2008.

DOORNIK, J. A.; HANSEN, H. A. An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, v. 70, p. 927-939, 2008.

DOWNING, D.; CLARK, J. *Estatística Aplicada*. 2. ed. São Paulo: Saraiva, 2005.

ENDERS, W. *Aplied Econometric Time Series*. 2. ed. New York: John Wiley & Sons, 2003.
FÁVERO, L. P. L. Dados em painel em contabilidade e finanças: teoria e aplicação. *Brazilian Business Review*, v. 10, n. 1, p. 131-156, 2013.
FÁVERO, L. P. L.; ALMEIDA, J. E. F. O comportamento dos índices de ações em países emergentes: uma análise com dados em painel e modelos hierárquicos. *Revista Brasileira de Estatística*, v. 72, n. 235, p. 97-137, 2011.
FÁVERO, L. P. L.; BELFIORE, P. Cash flow, earnings ratio and stock returns in emerging global regions: evidence from longitudinal data. *Global Economy and Finance Journal*, v. 4, n. 1, p. 32-43, 2011.
FÁVERO, L. P. L.; BELFIORE, P.; SILVA, F. L.; CHAN, B. L. *Análise de Dados*: Modelagem Multivariada para Tomada de Decisões. Rio de Janeiro: Elsevier, 2009.
FÁVERO, L. P. L.; SOTELINO, F. B. Elasticities of stock prices in emerging markets: a panel data approach. In: Batten, J. A.; Szilagyi, P. G. The Impact of the Global Financial Crisis on Emerging Financial Markets. *Contemporary Studies in Economic and Financial Analysis,* v. 93, p. 471-491, 2011.
FREES, E. W. *Longitudinal and Panel Data*: Analysis and Applications in the Social Sciences. Cambridge: Cambridge University Press, 2004.
GELMAN, A.; HILL, J. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Nova York: Cambridge University Press, 2007.
GIL, A. C. *Métodos e Técnicas de Pesquisa Social*. São Paulo: Atlas, 1999.
GREENE, W. H. *Econometric Analysis*. 6. ed. Upper Saddle River: Prentice Hall, 2007.
GUJARATI, D. N. *Econometria Básica*. 5. ed. Porto Alegre: Bookman, 2011.
HAMILTON, L. C. *Statistics with Stata*: Updated for Version 10. Belmont: Brooks/Cole, Cengage Learning, 2009.
HENRY, G. T. *Practical Sampling*. C. A.: Sage, 1990.
HILL, C.; GRIFFITHS, W.; JUDGE, G. *Econometria*. São Paulo: Saraiva, 2000.
HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. *Análise Exploratória de Dados:* Técnicas Robustas. Lisboa: Salamandra, 1983.
HOECHLE, D. Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal*, v. 7, n. 3, p. 281-312, 2007.
HOLTZ-EAKIN, D.; NEWEY, W.; ROSEN, H. S. Estimating vector autoregressions with panel data. *Econometrica*, v. 56, n. 6, p. 1371-1395, 1988.
HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. New York: John Wiley & Sons, 1989.
HSIAO, C. *Analysis of Panel Data*. 2. ed. Cambridge: Cambridge University Press, 2003.
ISLAM, N. Growth empirics: a panel data approach. *The Quarterly Journal of Economics*, v. 110, n. 4, p. 1127-1170, 1995.
JENKINS, S. P. Survival Analysis. Disponível em: http://michau.nazwa.pl/aska/uploads/Studenci/mag7_1.pdf (2005). Acesso em: 05/04/2013.
JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. New Jersey, Upper Saddle River: Prentice Hall, 1998.
JOHNSTON, J.; DINARDO, J. *Métodos Econométricos*. 4. ed. Lisboa: McGraw-Hill, 2001.
JONES, D. C.; KALMI, P.; MÄKINEN, M. The productivity effects of stock option schemes: evidence from Finnish panel data. *Journal of Productivity Analysis*, v. 33, n. 1, p. 67-80, 2010.
KACHIGAN, S. *Statistical Analysis*: An Interdisciplinary Introduction to Univariate & Multivariate Methods. New York: Radius Press, 1986.
KING, G.; KEOHANE, R. O.; VERBA, S. *Designing Social Inquiry*: Scientific Inference in Qualitative Research. Princeton: Princeton University Press, 1994.
KMENTA, J. *Elementos de Econometria*. São Paulo: Atlas, 1978.

KRISHNAKUMAR, J.; RONCHETTI, E. (org.). *Panel Data Econometrics*: Future Directions. Amsterdam: North Holland, 2000.
KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *Applied Linear Statistical Models*. 5. ed. New York: McGraw-Hill, 2004.
LEININGER, W. E. *Quantitative Methods in Accounting*. D. Van Nostrand Company, 1980.
LEVIN, R. I.; RUBIN, D. S. *Statistics for Management*. New Jersey: Prentice Hall, 1998.
LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. *Estatística*: Teoria e Aplicações. Rio de Janeiro: LTC, 2000.
LOHR, S. *Sampling*: Design and Analysis. New York: Duxbury, 1999.
LONG, S. J.; FREESE, J. *Regression Models for Categorical Dependent Variables*. Texas: Stata Corporation, 2001.
MADDALA, G. S. *Introdução à Econometria*. 3. ed. Rio de Janeiro: LTC, 2003.
MALHOTRA, N. K. *Pesquisa de Marketing*: Uma Orientação Aplicada. 3. ed. Porto Alegre: Bookman, 2001.
MAROCO, J. *Análise Estatística com Utilização do SPSS*. 5. ed. Lisboa: Silabo, 2011.
MARQUES, L. D. Modelos dinâmicos com dados em painel: revisão da literatura. Série Working Papers do Centro de Estudos Macroeconômicos e Previsão (CEMPRE) da Faculdade de Economia do Porto, Portugal, nº 100, 2000.
MARTINS, G. A. *Estatística Geral e Aplicada*. São Paulo: Atlas, 2001.
MATOS, O. C. *Econometria Básica*. São Paulo: Atlas, 1997.
MÁTYÁS, L.; SEVESTRE, P. (org.). *The Econometrics of Panel Data*: Fundamentals and Recent Developments in Theory and Practice. 3. ed. New York: Springer Verlag, 2008.
MEPHAM, M. *Accounting Models*. Polytech Publishers, 1980.
MILLS, T. C. *The Econometric Modelling of Financial Time Series*. Cambridge University Press, 1993.
MORETTIN, L. G. *Estatística Básica*: Probabilidade e Inferência. São Paulo: Pearson, 2009.
MORETTIN, P. A.; BUSSAB, W. O. *Estatística Básica*. 6. ed. São Paulo: Saraiva, 2010.
NEUFELD, J. L. *Estatística Aplicada à Administração Usando Excel*. São Paulo: Prentice Hall, 2003.
NEWBOLD, P. *Statistics for Business & Economics*. Prentice Hall, 1995.
PEVALIN, D.; ROBSON, K. *The Stata Survival Manual*. Maidenhead: Open University Press, 2009.
PINDYCK, R. S.; RUBINFELD, D. L. *Econometria:* Modelos e Previsões. 4. ed. Rio de Janeiro: Elsevier, 2004.
RABE-HESKETH, S.; EVERITT, B. *A Handbook of Statistical Analyses Using Stata*. 2. ed. Boca Raton: Chapman & Hall, 2000.
RAPPAPORT, A. *Information for Decision Making*. Prentice-Hall, 1982.
SHAPIRO, S.; FRANCIA, R. An approximation analysis of variance test for normality. *Journal of the American Statistical Association*, v. 67, p. 215-216, 1972.
SOARES, I. G.; CASTELAR, I. *Econometria Aplicada com o Uso do EViews*. Fortaleza: UFC/CAEN, 2003.
STATACORP. *Getting Started with Stata for Windows:* Version 11. College Station: StataCorp LP, 2009.
STATACORP. *Stata Statistical Software:* Release 12. College Station: StataCorp LP, 2011.
STEVENSON, W. J. *Estatística Aplicada a Administração*. São Paulo: Harbra, 1981.
STOCK, J. H.; WATSON, M. W. *Introduction to Econometrics*. 2. ed. Boston: Pearson Addison Wesley, 2007.
TABACHNICK, B. G.; FIDELL, L. S. *Using Multivariate Statistics*. USA: Allyn and Bacon, 2001.
TACQ, J. *Multivariate Analysis Techniques in Social Science Research*. Thousand Oaks: Sage Publications, 1996.
TAKAMATSU, R. T.; FÁVERO, L. P. L. Accruals, persistence of profits and stock returns in Brazilian public companies. *Modern Economy*, v. 4, p. 109-118, 2013.

VOGELVANG, B. *Econometrics*: Theory and Applications with EViews. Harlow: Financial Times Prentice Hall, 2005.
WEBSTER, A. *Estatística Aplicada à Administração e Economia*. São Paulo: McGraw-Hill, 2006.
WEISBERG, S. *Applied Linear Regression*. New York: John Wiley & Sons, 1985.
WONNACOTT, T.; WONNACOTT, R. J. *Introductory Statistics for Business and Economics*. 4. ed. New York: John Wiley & Sons, 1990.
WOOLDRIDGE, J. M. *Econometric Analysis of Cross Section and Panel Data*. 2. ed. Cambridge: MIT Press, 2010.
WOOLDRIDGE, J. M. *Introdução à Econometria*: Uma Abordagem Moderna. 4. ed. São Paulo: Cengage Learning, 2011.

ÍNDICE REMISSIVO